A UNIFIED SCALING LAW FOR TIME SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a Unified Scaling Law and a Unified Theory of Time Series Learning to understand how model capacity and historical information support forecasting. Across different lookback lengths and forecast horizons, we analyze 18,768 experimental cells from 21 checkpoints on 23 dataset-frequency tasks spanning six domains. Our empirical methodology integrates local resource relations into a parsimonious, fitted five-parameter law: capacity gains increase with history, context gains diminish toward saturation, and horizon effects enter as a common shift. Fitted without Toto 2.0, the law predicts its horizon-averaged capacity-scaling curves with mean absolute percentage errors of 1.09% and 1.50% at input lengths 2048 and 4096. To understand how history supports prediction, our learning theory uses Gaussian regression to analyze rule identification and predictive capability. We hypothesize that full-shot models learn by *accumulating information in weights*, while frozen time series foundation models (TSFMs) use history by *extracting information through activations*. Matched-history comparisons establish the predictive value of additional history. Controlled parameter exchanges and activation interventions provide evidence that history-derived rule information can be retained, reused across queries, and used to recover a contribution to long-context prediction. Together, these findings inform capacity scaling, context allocation, and the development of models that retain and apply historical rules.

1 INTRODUCTION

Time series foundation models (TSFMs), large neural models pretrained from scratch on time series, are becoming more capable as forecasting research expands (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024; Khwaja et al., 2026; Liu et al., 2026b). Public benchmarks such as GIFT-Eval and fev-bench bring together a growing range of forecasting models (Aksu et al., 2024; Shchur et al., 2026), while application studies investigate their use in financial forecasting, precipitation prediction, and battery degradation (Rahimikia et al., 2025; Song et al., 2026; Zhang et al., 2026). This progress has intensified debate over the value and scalability of foundation forecasting. Recent critiques question the premise of domain universality (Dai et al., 2026a). Empirical studies show that selected supervised baselines can compete with TSFMs under particular tasks and history budgets (Xu et al., 2025; Zhang & Gilpin, 2026). Alongside these critiques, early scaling studies establish benefits from increasing training resources and identify interactions between model complexity and history (Shi et al., 2024; Yao et al., 2025; Edwards et al., 2024). This raises the question: **under what resource and forecasting conditions does pretrained scale translate into useful predictive gains?** Existing work establishes training-resource scaling (Yao et al., 2025; Edwards et al., 2024), characterizes interactions of capacity and lookback (Shi et al., 2024), and studies how longer contexts support forecasting (Liu et al., 2025b; Butera et al., 2026). However, these works have not fully resolved the ongoing debate because their experiments remain limited in either scale (model and dataset diversity) or scope (pretrained capacity, available history, and forecasting horizon). This debate calls for more general and scalable quantification across heterogeneous TSFMs to analyze error under a wide range of tasks.

Therefore, we propose a **Unified Scaling Law**, aiming to deepen understanding of how TSFMs turn pretrained capacity and history information into useful forecasts, and to inform their future training and application. We fitted the Law through controlled experiments. The protocol yields 18,768 dataset–configuration–checkpoint cells and 1,070 aggregated response points from 21 checkpoints

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.

on 23 dataset-frequency tasks across six domains. Public results from GIFT-Eval provide cross-source evaluation on the same task panel. Our empirical methodology uses local resource relations and Occam's razor to guide law construction, parameter estimation, and predictive validation. As a case study, we withhold the Toto 2.0 series of five same-architecture checkpoints, spanning approximately 4M to 2.5B parameters, and use the law estimated from the remaining heterogeneous checkpoints to predict its capacity-scaling response. The Mean Absolute Percentage Errors (MAPEs) of the horizon-averaged capacity curves are 1.09% and 1.50% with input lengths of 2048 and 4096, showing the Law remains informative when applied to a held-out architecture.

Further, to understand why and how TSFMs work and how they relate to full-shot models, we propose a **Unified Theory of Time Series Learning**. History carries information about both the current state and the rule governing its evolution. We hypothesize that models trained on the target series (full-shot models) learn by *accumulating information in weights*, while frozen TSFMs use history by *extracting information through activations*, inspired by earlier work (Russin et al., 2025; Zhang & Gilpin, 2025; Xu et al., 2025). We analyze this learning problem through Gaussian regression and investigate the hypotheses through experiments matching the full-shot training-record length to the TSFM input length, complemented by controlled tests of rule identification, parameter retention, and cross-query activation transfer. The empirical methodology and learning theory provide a reusable basis for future scaling research as models and forecasting conditions evolve.

Our main contributions are:

- **A unified scaling law and empirical research methodology.** We develop a Unified Scaling Law linking capacity, input length, and output horizon through extensive experiments. We contribute an empirical methodology from local resource analysis to law construction, estimation, and validation on external public results and a held-out case study.
- **A unified theory of time series learning.** We combine Gaussian regression analysis with hypotheses about how TSFMs and full-shot models turn historical evidence into predictive rules. Matched-history comparisons and controlled parameter and activation interventions examine how rule information is retained and applied to new queries.
- **Guidance for model builders and users.** We use the Unified Scaling Law and learning analysis to give model developers empirical support for pretraining larger models. We guide users in selecting models, input lengths, and adaptation strategies for their forecasting requirements and deployment constraints.

2 Related work

Towards Time Series Foundation Models. Forecasting has progressed from statistical methods to Transformer and linear models (Nie et al., 2023; Zeng et al., 2023; Dai et al., 2026b), and to TSFMs pretrained across datasets (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024; Khwaja et al., 2026). Benchmarks compare these models (Aksu et al., 2024; Shchur et al., 2026), alongside debate over domain universality (Dai et al., 2026a). Comparisons with supervised models and history-based predictors highlight the roles of available history and adaptation (Xu et al., 2025; Zhang & Gilpin, 2025; Zhang et al., 2025; Zhang & Gilpin, 2026). Studies of in-weight learning and in-context inference offer a perspective on information use (Russin et al., 2025; Faw et al., 2025). We connect learning hypotheses with matched-history experiments and the Unified Scaling Law to study how TSFMs use forecasting resources. Our cross-query interventions examine whether historical rule information remains useful on independent queries.

Scaling laws for language and time series models. Language-model scaling laws relate loss to parameters, data, and computation (Kaplan et al., 2020; Hoffmann et al., 2022), with test-time scaling extending resource allocation to inference (Snell et al., 2025). Time series studies examine training-resource scaling and capacity–lookback relations (Shi et al., 2024; Yao et al., 2025; Edwards et al., 2024), while others study predictability and history requirements (Butera et al., 2026; Wang et al., 2026; Huang et al., 2026). Timer-XL, Toto 2.0, and YingLong investigate longer contexts, capacity-dependent horizon stability, and output scaling, respectively (Liu et al., 2025b; Khwaja et al., 2026; Wang et al., 2025). These studies establish resource effects in different settings but leave their joint description across heterogeneous released checkpoints unresolved. Our Unified Scaling Law connects pretrained capacity, input length, and the final scored horizon. Detailed related works are provided in Appendix A.

Table 1: Grid-first selection retains dataset-level results (cells); points aggregate responses for fitting and prediction. Controlled GIFT-Horizon contains only cells absent from Controlled Grid-Horizon.

Source	TSFM Checkpoints	Dataset-Frequency Task	Experimental Cells	Points
Controlled GIFT-Horizon	21	23	3,128	390
Controlled Grid-Horizon	21	23	15,640	680
Combined Controlled	21	23	18,768	1,070
GIFT Public	37	23	2,257	199

3 Problem definition and evidence design

We select a fixed panel $\mathcal{D}$ of 23 dataset-frequency pairs from GIFT-Eval based on available history and coverage across six domains, spanning 15 base datasets. Eligible evaluation cells provide more than 1,000 pre-origin observations. Appendix B lists the tasks and coverage.

Mean absolute scaled error (MASE) (Hyndman & Koehler, 2006) is a principal point-forecasting metric in GIFT-Eval (Aksu et al., 2024). Let m identify a checkpoint, $d \in \mathcal{D}$ a dataset-frequency pair, N the checkpoint's active parameter count in millions, L the available input length supplied to the model, and H the number of scored future steps. We define the normalized response R as raw MASE divided by that of Seasonal Naive (SN) at matching forecast origins and horizons:

$$R_{m,d,H,L} = \frac{\mathrm{MASE}^{\mathrm{raw}}_{m,d,H,L}}{\mathrm{MASE}^{\mathrm{raw}}_{\mathrm{SN},d,H}}. \tag{1}$$

Throughout the paper, figures, and tables, *MASE* denotes this normalized quantity R; values below one indicate lower error than SN. This dimensionless reference makes equal-weight averages interpretable across datasets with different units and magnitudes. Raw MASE uses the benchmark's seasonal scale and each forecast's median. Continuous calculations use data-supported input lengths; discrete displays group them into nominal levels, labelled simply as input length (Appendix B.4). We call (N, H, L) NHL coordinates and use natural logarithms.

The Controlled Grid-Horizon Panel varies input length and horizon at shared forecast origins; the Controlled GIFT-Horizon Panel retains native benchmark horizons and origins. Grid observations take precedence at overlapping conditions. Equal dataset means per checkpoint and resource condition, merged at repeated actual NHL coordinates within that checkpoint, yield 1,070 Controlled fitting points (Table 1). Removing overlapping checkpoint and resource bins from GIFT Public leaves 111 test points from 22 checkpoints. Appendix B specifies grids, aggregation, and overlap removal.

4 Deriving the Unified Scaling Law

Let $R(N, H, L)$ denote the aggregate response surface, and define log resources $n = \log N$, $h = \log H$, and $\ell = \log L$. The log-response surface is $f(n, h, \ell) = \log R(e^n, e^h, e^\ell)$.

Local resource relations. We summarize capacity contrasts at common horizon and context conditions across heterogeneous checkpoints. Within a checkpoint and horizon, adjacent input lengths provide context contrasts, represented by a first-order dependence of context elasticity on log length. Shared-origin horizon changes estimate horizon elasticity at fixed checkpoint and input. Let β denote the capacity–history coupling coefficient, γ the context coefficient, q the context-curvature coefficient, and α the horizon elasticity. These local approximations give

$$\frac{\partial f}{\partial n} = -\beta\ell, \tag{2}$$

$$\frac{\partial f}{\partial \ell} = -\gamma - \beta n + 2q\ell, \tag{3}$$

$$\frac{\partial f}{\partial h} = \alpha. \tag{4}$$

For $\beta > 0$, capacity gains increase with log history length; $q > 0$ represents diminishing context gains, with consistent mixed derivative $-\beta$. An affine extension adds a capacity-elasticity intercept

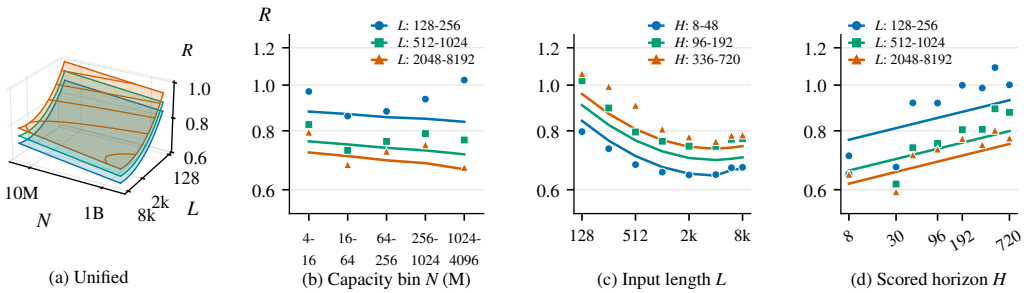

Figure 1: The Unified Scaling Law and observed resource profiles. (a) Predicted MASE over N, L at $H = 48, 192, 720$ (blue, green, orange). (b)–(d) Observed means (markers) and mean predictions (lines) over identical capacity, input-length, and horizon groups, using all 1,070 Controlled points. Resource axes are logarithmic; R is linear in (a) and logarithmic in (b)–(d). 1k denotes 1,024 steps.

λ, giving $\lambda - \beta\ell$ and an additional λn term after integration. Equation 2 selects the zero-intercept form for parsimony; Section 5.2 compares both forms and flexible alternatives. The horizon elasticity α describes changes relative to the matching baseline, whose error also depends on H. Appendix C.6 compares this common shift with nh and ℓh extensions.

Integration and the context saturation scale. Integrating these local approximations gives the five-parameter form with response intercept c (Appendix B.5):

$$\boxed{\log R = c + \alpha \log H - \gamma \log L - \beta \log N \log L + q(\log L)^2.} \tag{5}$$

For $q > 0$, define the operational *context saturation scale* $L_\star(N)$ as the input length where the fitted context elasticity reaches zero:

$$L_\star(N) = \exp\left(\frac{\gamma + \beta \log N}{2q}\right). \tag{6}$$

The fitted quadratic's stationary minimum in $\log L$ locates this scale: beyond this region, additional history yields little average improvement and may slightly increase error within the observed range through 8192 (Appendix B.2). Under the common-horizon-shift specification, H changes response height while N changes the saturation scale.

Estimation. We estimate the local relations before setting the response level. Capacity slopes at common horizon and context conditions are pooled across checkpoint subsets, and adjacent context contrasts hold checkpoint and horizon fixed. A robust objective fits (γ, β, q) to these targets over $\mathbb{R}^3$. We estimate α as the median matched horizon elasticity in the Controlled Grid-Horizon Panel and c as the median residual over the combined response points. Dataset-held-out and complete-series tests repeat every estimation step after excluding the held-out observations. Appendix Sections B.5 and B.6 specify the contrasts, weighting, and optimization.

5 Systematic Evidence for Joint Resource Scaling

5.1 Unified Response and Observed Resource Profiles

The combined controlled evidence gives the predicted response $\widehat{R}$:

$$\begin{aligned}\log \widehat{R} = {} & 0.965780 + 0.042866 \log H - 0.362138 \log L \\ & - 0.001846 \log N \log L + 0.022217 (\log L)^2.\end{aligned} \tag{7}$$

Figure 1 compares this prediction with observed capacity, context, and horizon profiles.

5.2 Selecting a Parsimonious Law for the Resource Relations

Part I: resource properties and reduced laws. The observed contrasts motivate three properties: capacity dependence, changing context returns, and capacity–history coupling. With subscripts on f

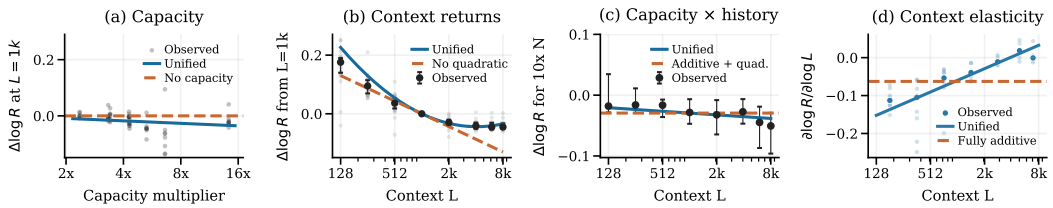

Figure 2: Observed resource relations and reduced-law restrictions: (a) capacity response, (b) changing context returns, (c) history-dependent capacity gains, and (d) context elasticity for the representative 22M group. Unified curves use fitted coefficients; orange dashed curves impose restrictions without refitting. Faint points show individual Controlled contrasts; emphasized markers show medians, with interquartile-range (IQR) bars where shown. Appendix C.4 specifies each comparison.

denoting partial derivatives, removing capacity or curvature forces $f_n = 0$ or $f_{\ell\ell} = 0$, respectively. Replacing the interaction with an additive capacity term forces $f_{n\ell} = 0$, making capacity gains independent of history and the context saturation scale independent of capacity. A fully additive form imposes both restrictions. Figure 2 compares these restrictions with Controlled observations; Appendix C.4 specifies contrasts; Appendix C.3 verifies the context trend on fixed complete support.

Part II: quantitative comparison under a parsimony prior. Occam's razor guides selection among forms that retain the observed properties: when predictive advantages are small, we prefer fewer parameters and fewer auxiliary shape assumptions. The product $n\ell$ is the lowest-degree interaction in log resources, and ℓ^2 is the lowest-degree term with nonzero context curvature, with no additional transition scale. A shallow minimum and a smooth plateau can express observed diminishing context gains through 8192. We re-estimate each candidate under the same dataset-held-out protocol to compare the Unified law with flexible alternatives and horizon-interaction extensions. Appendix Sections C.5 and C.6 give candidate forms, evaluation procedure, and supporting results.

5.3 Predictive validation

The selected law has 10.13% MAPE on 1,070 task-averaged Controlled response points. Applied directly to 111 non-overlapping GIFT Public points, it gives 12.82% MAPE, testing prediction under a different source of model results and horizon–context conditions (Figure 3a,b). Appendix C.1 gives reference comparisons and estimation stability.

Toto 2.0 provides five checkpoints spanning approximately 4M–2.5B parameters (Khwaja et al., 2026). We remove all five from both controlled sources and re-estimate the law using the remaining 16 heterogeneous checkpoints. Prediction on its 150 Controlled Grid-Horizon points gives 11.53% MAPE. After averaging over the five horizons at each capacity, the five-point capacity curves have MAPEs of 1.09% at $L = 2048$ and 1.50% at $L = 4096$ (Figure 3c,d). These predictions show that a statistical relationship identified across heterogeneous model designs can capture scaling within a held-out architecture. Appendix Figure 7 gives the complete Toto evaluation.

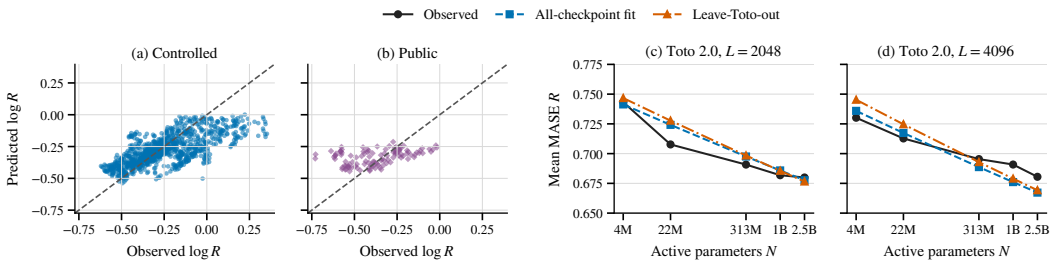

Figure 3: Controlled/Public prediction and horizon-averaged Toto 2.0 capacity curves.

6 A UNIFIED THEORY OF TIME SERIES LEARNING

Our Unified Theory of Time Series Learning studies how full-shot models and frozen TSFMs turn historical evidence into predictive rules. Historical evidence supports rule identification, the predictor's capabilities determine how accurately it estimates and applies the rule, and the learning path determines whether rule information is retained after training or extracted within the current forecast. Building on in-weight and in-context learning (Russin et al., 2025; Sander et al., 2024), we combine a regression analysis with two learning hypotheses and experiments on rule identification, retention, and reuse across queries.

6.1 HISTORICAL EVIDENCE AND THE TWO LEARNING PATHS

Let x_t denote the observation at time t, and $x_{u:v}$ the observations from indices u through v, inclusive. For query length ℓ_0, define history $D_{L,t} = x_{t-L+1:t}$ at forecast origin t and current query $B_t = x_{t-\ell_0+1:t}$. For a first-order autoregressive process, denoted AR(1), with coefficient ϕ, the next conditional mean is ϕx_t. Although the current state is x_t, estimating ϕ can require many observations, separating rule identification from current-state information (Butera et al., 2026).

Full-shot hypothesis: accumulating information in weights. We hypothesize that training can turn history information into rule-specific parameters that remain useful across subsequent queries. With training procedure $\mathcal{A}$ and forecasting model g, write the fitted parameters $\widehat{\theta}_L$ and forecast $\widehat{y}_{\text{weights}}$ as

$$\widehat{\theta}_L = \mathcal{A}(D_{L,t}), \qquad \widehat{y}_{\text{weights}} = g_{\widehat{\theta}_L}(B_t, H). \tag{8}$$

TSFM hypothesis: extracting information through activations. We hypothesize that a frozen TSFM can use historical evidence to form rule-specific activation states that its subsequent computation applies to the current query. Let F denote the TSFM, Θ_N its fixed pretrained parameters at capacity N, and $\widehat{y}_{\text{act}}$ its forecast:

$$\widehat{y}_{\text{act}} = F_{\Theta_N}(D_{L,t}, H), \qquad \Theta_N \text{ fixed.} \tag{9}$$

For a fixed query B, let p be the number of rule coefficients, indexed by $j = 1, \ldots, p$, with fixed features $\psi_j(B)$ and coefficient prior variances τ_j^2. Writing $\mathcal{N}(\mu, v)$ for a Gaussian distribution with mean μ and variance v, take independent rule coefficients $\omega_j \sim \mathcal{N}(0, \tau_j^2)$ and independent innovation ε with variance σ^2, distributed as $\mathcal{N}(0, \sigma^2)$. The scalar target is $Y = \sum_{j=1}^{p} \psi_j(B)\omega_j + \varepsilon$. With per-observation evidence precision $\nu_j > 0$, history supplies independent Gaussian observations of these coefficients with noise precision $L\nu_j$. A predictor uses posterior means for the first $r \leq p$ available coefficients and prior means for the remainder. Gaussian conditioning (Rasmussen & Williams, 2006) gives its prediction-error variance $V_r(L)$ and expected absolute error $\mathcal{E}_r(L)$:

$$V_r(L) = \sigma^2 + \sum_{j \leq r} \frac{\psi_j(B)^2 \tau_j^2}{1 + L\nu_j\tau_j^2} + \sum_{j > r} \psi_j(B)^2 \tau_j^2, \qquad \mathcal{E}_r(L) = \sqrt{\frac{2V_r(L)}{\pi}}. \tag{10}$$

It separates uncertainty reducible by history from predictive variation unused by this predictor.

In this model, the benefit of an additional useful coefficient increases with history length: more evidence improves its estimate. With positive remaining variance, the gains from further history diminish at sufficiently long inputs. The same rule estimate can be retained after training or computed from current history by a frozen predictor. For a single coefficient, training that leaves less uncertainty reduces the gain from additional immediate evidence about the same rule. Enlarging the available coefficient set creates opportunities to use history, while refining an existing estimate reduces the need for further evidence about it. Appendix D gives the assumptions and proofs.

6.2 PREDICTIVE VALUE OF TRAINING HISTORY AND IMMEDIATE INPUT

We compare DLinear and PatchTST (Zeng et al., 2023; Nie et al., 2023) with frozen TSFMs on 23 tasks at $L \in \{1024, 2048, 4096, 8192\}$ and $H \in \{96, 192, 336\}$. Full-shot models train on pre-origin history with local input $\ell_0 = 96$ and retain fitted states during scoring; TSFMs receive

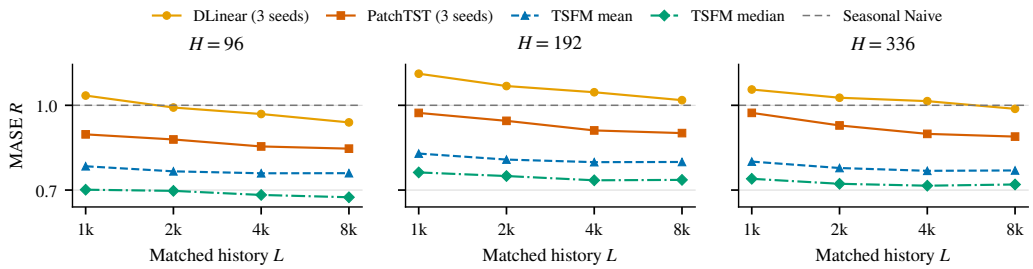

Figure 4: Forecasting with matched history budgets on 23 tasks. Full-shot curves average three seeds; TSFM curves show the mean and median over a fixed set of available checkpoints covering the grid. Seed error bars are too small to distinguish at this scale: mean pointwise relative standard deviations are $\pm 0.208\%$ for DLinear and $\pm 0.502\%$ for PatchTST. The line marks Seasonal Naive.

length-L histories at the same origins. This varies the shared history budget in earlier comparisons (Zhang & Gilpin, 2025; Xu et al., 2025); Appendix E.1 details implementation and training. Both full-shot architectures improve task-averaged MASE at every tested history extension and horizon (Figure 4). At $H = 96$, increasing history from 1,024 to 8,192 reduces DLinear's MASE from 1.042 to 0.931 and PatchTST's from 0.885 to 0.833. The fixed TSFM panel's mean and median remain lower at every matched condition; PatchTST's gaps to both narrow over this history range at all three horizons.

A complementary crossed experiment varies training history $T \in \{4096, 8192\}$ and immediate input $C \in \{96, 512\}$ independently at $H = 96$, on the same 23 tasks with three seeds. Alongside DLinear and two PatchTST sizes, we compare the same TimesFM 2.5 checkpoint before and after task-specific adaptation, with each input configuration trained and validation-selected separately. MASE falls from 0.9233 to 0.7877 under adaptation, a 14.68% reduction in the task-averaged score. Appendix E.1 reports the results. The MASE gap between the 96-step and 512-step TimesFM configurations decreases from 0.2396 before adaptation to 0.0768 afterwards, a reduction of about 68%. This supports partial substitution between retaining information through training and supplying longer immediate input, as illustrated by the single-rule analysis in Appendix D.3.

6.3 Rule identification and retention

We use three stationary processes: Gaussian AR(1), Gaussian lag-8, and a nonlinear threshold process. Within each process, two rule signs $s \in \{-1, +1\}$ share a single-observation distribution with mean zero and variance one. The rule changes the temporal transition relation. For each test query, the last 96 observations are identical under both rules; conditionally generated earlier histories provide evidence for the alternative transitions. Independent queries include both signs and multiple amplitudes, so applying a rule requires combining historical information with the new query.

We evaluate one-step predictions against the known conditional mean using *conditional-mean MASE*, measuring rule agreement without future innovation noise. The naive reference uses the last query observation and the same target. Appendix E.2 specifies the processes, sampling, scoring, and independent evidence splits.

Historical evidence and rule identification. TimesFM 2.5 and Chronos 2 improve conditional-mean MASE from 96 to 2,048 observations, with positive paired-query 95% bootstrap intervals in all 36 model–process–magnitude–normalization conditions. Fixed population statistics preserve the gains. An oracle knowing the process family and magnitude, and inferring only the sign from the historical prefix, reaches conditional-mean MASE below 0.001 at rule magnitude $a = 0.15$ and $L = 512$ for all three processes (Appendix E.3). With queries and marginal distributions fixed, historical transition structure permits accurate rule identification, while TSFM errors and rule responses continue to vary with input length.

Trained parameters retain rules for new queries. We train DLinear and a compact PatchTST separately under each rule using three history lengths and three seeds, for 108 fits. Each validation-

Table 2: Shared-site transfer and recovery in TimesFM 2.5. Conditional-mean MASE, averaged over three rule magnitudes and two donor repetitions. Natural, erased, and recovered predictions share the 512-step recipient history.

Process	Query only	Transfer	Natural	Erased	Recovered
AR(1)	0.26289	0.24181	0.17697	0.23986	0.21229
Lag-8	0.26972	0.25106	0.19702	0.25074	0.23755
Threshold	0.30168	0.27873	0.24983	0.28404	0.26195

selected model then predicts independent queries using only their 96 current observations. Parameters are exchanged between opposite-rule models, with normalization buffers calibrated on a shared independent history mixture. Predictions move toward the donor rule in all 54 process–architecture–length–seed conditions. Both architectures also improve their mean forecast accuracy as training history increases. At $L = 8192$, both architectures attain conditional-mean MASE below 0.13 on the Gaussian processes, averaged over three seeds (Appendix E.4).

6.4 Rule transfer and recovery through activations

We intervene inside frozen TimesFM 2.5 using activation patching (Zhang & Nanda, 2024). We average four independent donors' intermediate-state differences between 512 observations and their final 96 observations. The primary layer, token span, history length, and strength are fixed across processes and experiments. For *transfer*, we add this historical state difference to the short-input state of a new query. For *recovery*, we replace the corresponding state in the query's natural long-context execution with its short-input state, then add the same donor difference; other long-context states remain unchanged. Transfer tests cross-query reuse; recovery tests restoration of its contribution to natural long-context prediction.

Transfer reduces mean error by 7.51%; independent-donor recovery restores 41.67% of the mean error increase caused by erasure at the selected site. Erasure raises error, and transfer and recovery improve it with positive paired-query 95% intervals in all 18 conditions. Both interventions outperform random and carrier-only controls; rule-dependent gains also hold against shuffled-history donors (Appendix E.5). With recipient query and history fixed, matching-rule donors outperform opposite-rule donors for both interventions in every tested TimesFM condition across three layers and three strengths. Across the nine layer–strength settings, mean recovery improves over erasure in all nine and mean direct transfer over query-only prediction in seven. Chronos 2 shows the same matching advantage on independent confirmation data at a development-selected setting: recovery lowers mean error by 6.01%, with positive intervals in all 18 conditions, while direct transfer leaves overall error nearly unchanged. Appendix E.5 reports both models' controls and intervention robustness.

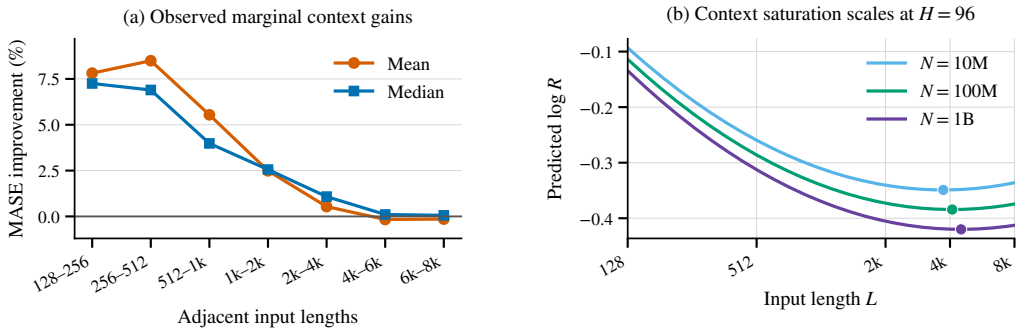

Figure 5: Context gains and fitted saturation scales. Left: mean and median MASE improvements between adjacent inputs for matched checkpoint–horizon pairs. Right: fitted $\log R$ at $H = 96$ for three capacities; $L_\star$ markers locate approximate saturation regions within the evaluated range.

7 PRACTICAL GUIDANCE FOR MODEL BUILDERS AND USERS

7.1 FOR MODEL BUILDERS: COORDINATE CAPACITY, CONTEXT, AND HORIZON

The Unified Scaling Law links capacity, input length, and forecast horizon, providing a basis for joint model and training design. For the fitted $\beta, q > 0$, capacity gains at fixed horizon and input length and the capacity dependence of the saturation scale satisfy

$$\frac{\partial \log R}{\partial \log N} = -\beta \log L, \qquad \frac{d \log L_\star}{d \log N} = \frac{\beta}{2q} > 0. \tag{11}$$

Increasing capacity lowers fitted error and shifts saturation toward longer histories, providing empirical support for pretraining larger models. Because capacity gains depend on history, compare model sizes across the input lengths intended for deployment. Training and validation should cover the application's input–horizon combinations. The horizon term describes performance relative to the matched baseline, so validation should retain horizon-specific references when assessing these gains. Alongside average error, evaluate how models use historical information. The controlled tests in Section 6 assess whether historical rules are identified, retained, and applied across queries. An oracle can distinguish insufficient historical evidence from inaccurate use of available evidence; measuring rule responses together with forecast error helps identify improvements in history use.

7.2 FOR MODEL USERS: SELECT MODELS AND ALLOCATE HISTORY

Start from the required forecast horizon and available history, then compare model capacity and input length jointly. The fitted capacity relation and Toto 2.0 observations support considering larger checkpoints, particularly when longer histories are available (Khwaja et al., 2026). For fixed capacity and horizon, the law gives

$$\frac{\partial \log R}{\partial \log L} = 2q \log \frac{L}{L_\star(N)}. \tag{12}$$

Context gains weaken near the fitted saturation scale, providing a broad starting region for task-level validation (Figure 5). Evaluate candidate models and contexts at the required horizon under the deployment latency and memory constraints; when errors are similar, shorter inputs reduce both costs. Seasonal Naive provides a reference for deciding when a TSFM offers useful gains, including conditions with short histories or long horizons. A frozen TSFM is a natural starting point when task-specific training is unavailable; full-shot training can retain information from a long record for repeated short-input forecasts. The crossed experiment motivates validating adaptation together with deployment input length. Retune context after adaptation using the adapted model's own validation curve: information learned during training can make shorter-input configurations more competitive.

8 DISCUSSION AND CONCLUSION

Discussion. The Unified Scaling Law describes average resource returns, while our learning analysis distinguishes evidence available in history from its predictive use. Retention and extraction provide complementary ways to apply historical rules. Task-specific adaptation can shift the useful balance between retained information and immediate input. These findings motivate evaluating history-use capability alongside capacity and jointly validating training and deployment choices.

Conclusion. We establish a five-parameter Unified Scaling Law through observed resource properties, parsimony, and validation on Public results and held-out Toto 2.0 checkpoints. Our Unified Theory of Time Series Learning combines Gaussian regression with matched-history comparisons and controlled interventions. Parameter exchanges support retained rule information, while cross-query activation transfer and recovery support its reuse in frozen prediction. The empirical methodology provides a reusable sequence from local resource analysis to estimation and validation. Together with the learning theory, it supports future scaling studies as models and forecasting settings evolve, while informing capacity scaling, context allocation, and model development.

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

We release the core code needed to reproduce our experiments. Since evaluating a broad range of time series foundation models requires complex environment configurations and substantial computational resources, we also include the main evaluation results to facilitate further research by the community. The code and results are included in the supplementary material submitted with this paper.

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. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

Miquel Noguer I Alonso and Rodolfo Pereira Franklin. Pretrained time-series foundation models for financial return forecasting. *arXiv preprint arXiv:2606.27100*, 2026.

Abdul Fatir Ansari, Lorenzo Stella, 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 Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

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.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 57529–57580. Curran Associates, Inc., 2025. doi: 10.52202/085713-1929.

Luca Butera, Giovanni De Felice, Andrea Cini, and Cesare Alippi. Why do time series models need long context windows? *arXiv preprint arXiv:2606.01999*, 2026.

Sanjay Chakraborty, Ibrahim Delibasoglu, and Fredrik Heintz. Scaling transformers for time series forecasting: do pretrained large models outperform small-scale alternatives? *Artificial Intelligence Review*, 59(2):62, January 2026. ISSN 1573-7462. doi: 10.1007/s10462-025-11481-7.

Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean du Terrail, Anna-Monica Toon, Kan Wang, Stephan Xie, Zongzhe Xu, Viktoriya Zhukova, David Asker, Ameet Talwalkar, and Othmane Abou-Amal. This time is different: An observability perspective on time series foundation models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 50907–50951. Curran Associates, Inc., 2025. doi: 10.52202/085713-1698.

Xilin Dai, Wanxu Cai, Zhijian Xu, and Qiang Xu. Position: Universal time series foundation models rest on a category error. *arXiv preprint arXiv:2602.05287*, 2026a.

Xilin Dai, Zhijian Xu, Wanxu Cai, and Qiang Xu. From samples to scenarios: A new paradigm for probabilistic forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. 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. 10148–10167. PMLR, 21–27 Jul 2024.

Thomas D. P. Edwards, James Alvey, Justin Alsing, Nam H. Nguyen, and Benjamin D. Wandelt. Scaling-laws for large time-series models. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

EINK. CleanTS model card. Hugging Face model card for CleanTS-65M, n.d. URL `https://huggingface.co/EINK/CleanTS-65M`. Accessed September 24, 2026.

Matthew Faw, Rajat Sen, Yichen Zhou, and Abhimanyu Das. In-context fine-tuning for time-series foundation models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16355–16374. PMLR, 13–19 Jul 2025.

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.

FlowVortex. LiteSpecFormer. GitHub repository, 2026. URL `https://github.com/FlowVortex/LiteSpecFormer`.

Xinghong Fu, Yanhong Li, Georgios Papaioannou, and Yoon Kim. Reverso: Efficient time series foundation models for zero-shot forecasting. In *Second ICML Workshop on Foundation Models for Structured Data*, 2026.

Lars Graf, Thomas Ortner, Stanisław Woźniak, and Angeliki Pantazi. FlowState: Sampling-rate-equivariant time-series forecasting. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Gredio Labs. goia-forecast-nano-v0. Hugging Face model card, n.d. URL `https://huggingface.co/gredio/goia-forecast-nano-v0`. Accessed September 24, 2026.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. doi: 10.52202/068431-2176.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: Extending tabPFN-v2 to time series forecasting. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Ruijin Hua, Zichuan Liu, Kun Zhang, and Yiyuan Yang. Diversified scaling inference in time series foundation models. *arXiv preprint arXiv:2601.17376*, 2026.

Lang Huang, Jinglue Xu, and Luke Darlow. How good can linear models be for time-series forecasting? *arXiv preprint arXiv:2606.27282*, 2026.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, October 2006. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2006.03.001.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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.

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*, 2026a.

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Junnan Li, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-MoE: Empowering time series foundation models with sparse mixture of experts. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38940–38962. PMLR, 2025a.

Yiding Liu, Yifan Hu, Hongjie Xia, Peiyuan Liu, Hongzhou Chen, Xilin Dai, Zewei Dong, and Jiang-Ming Yang. Falcon-X: A time series foundation model for heterogeneous multivariate modeling. *arXiv preprint arXiv:2605.27286*, 2026b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-xl: Long-context transformers for unified time series forecasting. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 83982–84006, 2025b.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317. PMLR, 2025c.

Yong Liu, Xingjian Su, Shiyu Wang, Haoran Zhang, Haixuan Liu, Yuxuan Wang, Zhou Ye, Yang Xiang, Jianmin Wang, and Mingsheng Long. Timer-S1: A billion-scale time series foundation model with serial scaling. *arXiv preprint arXiv:2603.04791*, 2026c.

Lucas Meyer, Claudio Sole, Huikan Xiang, Nicolas Li, Lucas Franceschino, Arnau Quera-Bofarull, Maarten P. Scholl, Joachim Fainberg, and Geoffrey Négiar. t_0: A time-series foundation model for forecasting with context. *arXiv preprint arXiv:2609.24559*, 2026a.

Marcel Meyer, Sascha Kaltenpoth, Kevin Zalipski, and Oliver Müller. Rethinking evaluation in the era of time series foundation models: (un)known information leakage challenges. *arXiv preprint arXiv:2510.13654*, 2026b.

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.

Marco Obermeier, Marco Pruckner, Florian Haselbeck, and Andreas Zeiselmair. FETS benchmark: Foundation models enable scalable and generalizable energy time series forecasting. *Energy and AI*, 26:100867, December 2026. ISSN 2666-5468. doi: 10.1016/j.egyai.2026.100867.

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.

Eghbal Rahimikia, Hao Ni, and Weiguan Wang. Re(visiting) time series foundation models in finance. *arXiv preprint arXiv:2511.18578*, 2025.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. ISBN 0-262-18253-X.

Jacob Russin, Ellie Pavlick, and Michael J. Frank. Parallel trade-offs in human cognition and neural networks: The dynamic interplay between in-context and in-weight learning. *Proceedings of the National Academy of Sciences*, 122(35):e2510270122, August 2025. ISSN 1091-6490. doi: 10.1073/pnas.2510270122.

Michael Eli Sander, Raja Giryes, Taiji Suzuki, Mathieu Blondel, and Gabriel Peyré. How do transformers perform in-context autoregressive learning ? 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. 43235–43254. PMLR, 21–27 Jul 2024.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. *arXiv preprint arXiv:2509.26468*, 2026.

Jingzhe Shi, Qinwei Ma, Huan Ma, and Lei Li. Scaling law for time series forecasting. 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. 83314–83344. Curran Associates, Inc., 2024. doi: 10.52202/079017-2650.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 34635–34667, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 10131–10165, 2025.

MinJi Song, YeonJi Chung, Ju-Young Shin, and Myoung-Jin Um. A comparative study on monthly precipitation forecasting using time-series foundation models in south korea. *Engineering Applications of Artificial Intelligence*, 179:115143, 2026. doi: 10.1016/j.engappai.2026.115143.

vilhess. A foundation model for univariate time series forecasting. GitHub repository, n.d. URL `https://github.com/vilhess/PatchFM`. Accessed September 26, 2026.

Xue Wang, Tian Zhou, Jinyang Gao, Bolin Ding, and Jingren Zhou. Output scaling: YingLong-delayed chain of thought in a large pretrained time series forecasting model. *arXiv preprint arXiv:2506.11029*, 2025.

Yuxuan Wang, Haixu Wu, Yuezhou Ma, Yuchen Fang, Ziyi Zhang, Yong Liu, Shiyu Wang, Zhou Ye, Yang Xiang, Jianmin Wang, and Mingsheng Long. Exploring accuracy law for deep time series forecasters: An empirical study. *arXiv preprint arXiv:2510.02729*, 2026.

Yunshi Wen, Wesley M. Gifford, Chandra Reddy, Lam M. Nguyen, Jayant Kalagnanam, and Anak Agung Julius. Revisiting the generic transformer: Deconstructing a strong baseline for time series foundation models. *arXiv preprint arXiv:2602.06909*, 2026.

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, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with $10k$ parameters. In *The Twelfth International Conference on Learning Representations*, 2024.

Zongzhe Xu, Ritvik Gupta, Wenduo Cheng, Alexander Shen, Junhong Shen, Ameet Talwalkar, and Mikhail Khodak. Specialized foundation models struggle to beat supervised baselines. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 102908–102933, 2025.

Siqiao Xue, Zhaoyang Zhu, Wei Zhang, Rongyao Cai, Rui Wang, Yixiang Mu, Fan Zhou, Jianguo Li, Peng Di, and Hang Yu. QuitoBench: A high-quality open time series forecasting benchmark. *arXiv preprint arXiv:2603.26017*, 2026.

Qingren Yao, Chao-Han Huck Yang, Renhe Jiang, Yuxuan Liang, Ming Jin, and Shirui Pan. Towards neural scaling laws for time series foundation models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 83793–83820, 2025.

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 transformers for time series: Rank structure, flow-of-ranks, and compressibility. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 42449–42489, 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.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024.

Lu Zhang, Demba Diallo, Claude Delpha, and Mohamed Benbouzid. Cross-dataset battery life forecasting with time-series foundation models: From zero-shot to prefix-adapted. *Energy and AI*, 25:100771, 2026. doi: 10.1016/j.egyai.2026.100771.

Yuanzhao Zhang and William Gilpin. Zero-shot forecasting of chaotic systems. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 93873–93899, 2025.

Yuanzhao Zhang and William Gilpin. Context parroting: A simple but tough-to-beat baseline for foundation models in scientific machine learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 58314–58340, 2026.

Yunkai Zhang, Qi Zeng, Yawen Zhang, Zhijie Xu, Ming Zheng, Chongyang Gao, Muyan Jiang, and Zeyu Zheng. Are time series foundation models ready for zero-shot forecasting? In *1st ICML Workshop on Foundation Models for Structured Data*, 2025.

A Detailed related works

A.1 Time series foundation models

Time series forecasting has expanded from statistical methods to Transformer-based models such as PatchTST and iTransformer (Nie et al., 2023; Liu et al., 2024) and linear models such as DLinear, FITS, and TimePrism (Zeng et al., 2023; Xu et al., 2024; Dai et al., 2026b), alongside the recent rise of large-scale pretraining (Kottapalli et al., 2025). Chronos casts probabilistic prediction as token modeling (Ansari et al., 2024), TimesFM develops a decoder-only forecaster (Das et al., 2024), and Moirai introduces unified training across datasets and frequencies (Woo et al., 2024). Subsequent releases include Chronos-Bolt and Chronos-2 (Ansari et al., 2024; 2025), Moirai-MoE and Moirai 2 (Liu et al., 2025a; 2026a), and the TiRex and TiRex-2 forecasting models (Auer et al., 2025; Podest et al., 2026). Other models include Sundial (Liu et al., 2025c), Granite FlowState (Graf et al., 2026), and Granite PatchTST-FM (Wen et al., 2026), alongside the Toto and Toto 2.0 series (Cohen et al., 2025; Khwaja et al., 2026). Time-MoE and Timer-S1 explore expert-based scaling (Shi et al., 2025; Liu et al., 2026c), while Timer-XL develops expanded forecasting contexts (Liu et al., 2025b); other released forecasters include Falcon-X, Kairos, YingLong, and Reverso (Liu et al., 2026b; Feng et al., 2026; Wang et al., 2025; Fu et al., 2026). The public comparison also covers TabPFN-TS (Hoo et al., 2026), and the forecasting releases t0, CleanTS, goia, PatchFM, and LiteSpecFormer (Meyer et al., 2026a; EINK, n.d.; Gredio Labs, n.d.; vilhess, n.d.; FlowVortex, 2026). GIFT-Eval and fev-bench provide comparisons across heterogeneous domains, frequencies, and forecast lengths (Aksu et al., 2024; Shchur et al., 2026).

A.2 Scaling laws for language and time series models

Language-model scaling laws characterize loss through parameters, training data, and computation (Kaplan et al., 2020; Hoffmann et al., 2022). Test-time scaling extends the allocation problem to inference (Snell et al., 2025). For time series, Shi et al. (2024) jointly analyze data size, model complexity, and lookback with a fixed forecast horizon, finding that useful lookback depends on the training-data regime. Yao et al. (2025) study parameter, data, and compute scaling across in-domain and out-of-domain forecasting, while Edwards et al. (2024) study training-resource laws with a fixed context. Complementary studies examine predictability and context selection. Butera et al. (2026) explain why identifying a generating process can require history longer than its memory, while Sander et al. (2024) analyze how Transformers estimate autoregressive dynamics from context. Wang et al. (2026) relate the lowest forecasting errors attained by deep models to window-wise pattern complexity, while Huang et al. (2026) show that the relationship between useful lookback and forecast horizon varies across datasets for tuned linear models.

Timer-XL examines longer contexts and the dependence of their benefits on model and training choices (Liu et al., 2025b). Toto 2.0 demonstrates improved forecasting performance with increasing model size and examines how capacity affects long-horizon stability (Khwaja et al., 2026). YingLong studies output scaling (Wang et al., 2025). Diversified inference investigates sampling and perturbations alongside capacity and context length (Hua et al., 2026). These findings concern different architectures and inference procedures, without establishing a shared error law across heterogeneous released checkpoints.

A.3 Analysis and controversy surrounding TSFMs

The growth of foundation forecasting has prompted scrutiny of its conceptual scope and evaluation (Dai et al., 2026a; Meyer et al., 2026b). Empirical comparisons examine the performance of pretrained forecasting models across model scales and application domains (Chakraborty et al., 2026; Alonso & Franklin, 2026). Xu et al. (2025) and Zhang & Gilpin (2025) compare frozen TSFMs with supervised baselines given the same history, using autoregressive models on forecasting benchmarks and N-BEATS on chaotic systems, respectively. Zhang et al. (2025) combine a representation learned by principal component analysis on a multi-domain corpus with a linear head trained on the available history, questioning the practical necessity of zero-shot prediction. Context parroting shows that matching and copying historical segments can rival TSFMs on dynamical-system forecasting and relates one-step error to history length (Zhang & Gilpin, 2026). QuitoBench and FETS report context- and horizon-dependent model comparisons on application-traffic and energy data,

respectively (Xue et al., 2026; Obermeier et al., 2026). These studies establish the importance of history budgets and task-specific adaptation.

An information-use perspective connects these comparisons to learning mechanisms. Russin et al. (2025) distinguish incremental in-weight learning from in-context inference through activation dynamics. In forecasting, Faw et al. (2025) investigate adaptation through examples supplied in context and through weight fine-tuning, while Butera et al. (2026) explain how training or inference supports identification of a generating process. Rank and compressibility analyses examine how TSFMs represent predictive information (Yu et al., 2026). We develop learning hypotheses about history information and examine them through matched-history experiments with linear and Transformer full-shot models (Zeng et al., 2023; Nie et al., 2023).

B Evaluation protocol and statistical estimation

B.1 Benchmark snapshot and common dataset scope

We use fixed GIFT-Eval benchmark, dataset, and public-result snapshots documented in the released code. Table 3 lists the 23 dataset-frequency pairs selected based on available history and domain coverage. They span 15 base datasets and six benchmark domains; separate frequencies constitute distinct forecasting tasks. Cloud and web operations include virtual-machine and operational monitoring, energy includes electricity demand, transformer measurements, and solar power, and environmental data include weather, air quality, and river flow. Transportation, retail sales, and births supply additional application domains. This fixed panel is shared by the controlled experiments, public analysis, Toto case study, and matched-history comparison, supplying 61 native GIFT configurations and supporting shared-origin evaluation through maximum horizon $H_{\max} = 720$.

Controlled evaluations use the same evaluator with experiment-specific input cropping and forecast-origin construction. The public-inclusion filter follows the registered checkpoint catalogue, while the controlled experiment uses the checkpoint panel in Table 4. Both analysis sources retain cells whose available history exceeds 1,000 observations; controlled eligibility uses the shortest available history in the cell. This condition concerns the data record, and allocated contexts can be as short as 128 steps.

Table 3: The fixed evaluation panel and its application domains. Sampling intervals span five, ten, and fifteen minutes, one hour, one day, and one week. Nineteen pairs have short, medium, and long GIFT configurations; the four daily/weekly pairs contribute four short configurations, for 61 in total.

Dataset	Sampling intervals	Domain
Bitbrains fast storage	5 min	Cloud/web operations
Bitbrains random	5 min	Cloud/web operations
BizITObs L2C	5 min, 1 h	Cloud/web operations
Electricity	15 min, 1 h	Energy
ETT1	15 min, 1 h	Energy
ETT2	15 min, 1 h	Energy
Hierarchical sales	1 day	Sales
Jena weather	10 min, 1 h	Environment
KDD Cup 2018	1 h	Environment
LOOP Seattle	5 min, 1 h	Transportation
M Dense	1 h	Transportation
Saugeen	1 day, 1 week	Environment
Solar	10 min, 1 h	Energy
SZ Taxi	15 min	Transportation
US births	1 day	Healthcare
15 base datasets	23 dataset-frequency pairs	6 domains

B.2 CHECKPOINT CAPACITIES AND INPUT COVERAGE

The controlled panel includes 21 checkpoints with heterogeneous architectures, pretraining choices, and inference interfaces. Table 4 records their active and total capacities; active capacity defines the N coordinate. Checkpoint names identify the released weights.

Table 4: Controlled checkpoints and active and total parameter counts, in millions.

Checkpoint	Active N	Total N	Checkpoint	Active N	Total N
Toto 2.0 4M	4.14	4.14	Moirai 1.1 large	310.97	310.97
Toto 2.0 22M	21.92	21.92	Moirai 2.0 small	11.39	11.39
Toto 2.0 313M	312.68	312.68	TimesFM 1.0	200.00	200.00
Toto 2.0 1B	1041.03	1041.03	TimesFM 2.0	500.00	500.00
Toto 2.0 2.5B	2454.28	2454.28	TimesFM 2.5	200.00	231.29
Chronos 2 small	27.93	27.93	TiRex 1.1 GIFT	35.00	35.00
Chronos 2	119.48	119.48	TiRex-2 GIFT-ZS	38.40	82.50
Chronos-Bolt small	47.72	47.72	Sundial base	128.33	128.33
Chronos-Bolt base	205.29	205.29	Granite FlowState r1.1	18.52	18.52
Moirai 1.0 small	13.83	13.83	Granite PatchTST-FM r1	257.90	257.90
Moirai 1.1 base	91.36	91.36			

The active-capacity convention follows each checkpoint's inference path. TiRex-2 uses 38.4M active parameters out of 82.5M total for univariate inference (Podest et al., 2026). TimesFM (Das et al., 2024) capacities follow the metadata of its 2.0 and 2.5 checkpoints, with the latter's additional quantile head included in the total count alongside its 200M backbone. Nominal capacities such as 200M and 500M preserve the precision available in the metadata. Frozen inference settings follow each checkpoint's benchmark profile; exact configurations, context limits, and snapshot provenance are provided in the released code.

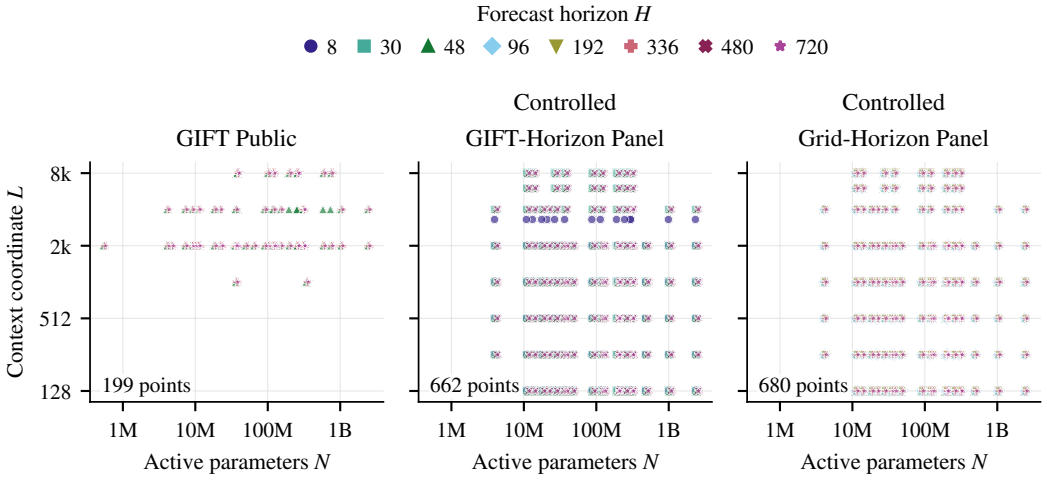

Figure 6: Resource coverage across active capacity N, input length L, and scored horizon H (colors and markers). Public coordinates are grid-binned; Controlled coordinates use data-supported inputs. The Controlled GIFT-Horizon Panel shows 662 protocol points and contributes 390 additional points after Grid-first selection (Table 1). Small horizontal offsets separate overlapping horizons.

Long-context support. The Controlled input grid ends at 8192, an allocation supported by nine of the 21 registered checkpoint interfaces. Under the shared-origin protocol, 12 of the 23 dataset-frequency tasks provide at least 8192 preceding observations at every scored window, and 13 provide them at one or more windows. Shorter histories are cropped per window, so an aggregated long-context coordinate can include datasets with shorter available histories. Appendix C.3 compares fixed task and checkpoint sets with complete recorded input support.

B.3 SHARED-ORIGIN HORIZON CONSTRUCTION

Both controlled protocols scan allocated input lengths $L \in \{128, 256, 512, 1024, 2048, 4096, 6144, 8192\}$ within each checkpoint's registered context limit. The Controlled GIFT-Horizon Panel uses the 61 official configurations, while the Controlled Grid-Horizon Panel scores $H \in \{48, 96, 192, 336, 720\}$ at shared origins across the same 23 dataset-frequency pairs. The additional GIFT horizons $H = 8, 30, 480$ cover one, three, and nineteen pairs, respectively.

For a benchmark configuration with shortest series length $n_{\min}$ and scored horizon H, GIFT's window count $W(H)$ is

$$W(H) = \min(20, \max(1, \lceil 0.1 n_{\min}/H \rceil)) \,. \tag{13}$$

Let n_i denote the length of series i and $t_{i,k}$ its cutoff for window k. The Controlled GIFT-Horizon Panel retains the official H, window count, and origin spacing, giving $t_{i,k} = n_i - W(H)H + kH$ for $k = 0, \ldots, W(H) - 1$. The Controlled Grid-Horizon Panel instead uses the common window count $W^* = W(720)$ and

$$t_{i,k} = n_i - W^* 720 + k720, \qquad k = 0, \ldots, W^* - 1. \tag{14}$$

Each horizon scores the first H observations after these shared cutoffs, with Seasonal Naive evaluated at the same horizon and origins. Full-shot evaluation reuses this construction for $H = 96, 192, 336$.

Input cropping is applied independently at each window, with historical observations and past-only covariates sharing the same time boundary. Each checkpoint retains its native forecasting path, including direct prediction, internal recursion, or masked-future generation; H denotes the final scored interval.

B.4 RESPONSE CONSTRUCTION AND DISPLAY COORDINATES

Cells are identified by checkpoint, dataset-frequency, scored horizon, and allocated input length. Before aggregation, Grid-first selection retains the Grid observation at each shared key, including its forecast origins and matched baseline, at overlapping horizons $H = 48, 720$. Retained dataset-cell responses are averaged equally within each checkpoint, horizon, and allocated input condition, using its data-supported long-end context. Repeated actual NHL coordinates are merged within each checkpoint; checkpoints with equal parameter counts retain separate identities. Logarithms are taken after response averaging, and final response points have equal weight in prediction-error summaries.

Continuous fitting and prediction use data-supported actual context lengths; predefined input-length levels organize discrete displays and matched local comparisons. Figure 1b groups capacity into five intervals bounded by $\{4, 16, 64, 256, 1024, 4096\}$M. The capacity and horizon panels group curves by input length, while the input-length panel groups curves by horizon. Each panel uses all Controlled response points: predictions are evaluated at individual actual NHL coordinates, then averaged with the same membership and weights as the observations.

Public configurations retain the original seven-level context grid $\{128, 256, 512, 1024, 2048, 4096, 8192\}$ and original horizon bins. For cross-source overlap alone, Controlled coordinates are mapped onto that same Public grid; shared checkpoint–horizon-bin–context-bin keys are excluded from Public evaluation.

B.5 EMPIRICAL RELATIONS AND LOCAL TARGETS

Suppose f is differentiable on a connected resource region and obeys Eqs. 2–4. With integration function $g(h, \ell)$, the capacity relation gives $f = -\beta n\ell + g(h, \ell)$. Substituting into the context relation yields $\partial g/\partial \ell = -\gamma + 2q\ell$. Introducing the remaining integration function $a(h)$ gives $g = -\gamma\ell + q\ell^2 + a(h)$. Finally, the horizon relation gives $a(h) = \alpha h + c$, recovering Eq. 5.

Before constructing local contrasts, points sharing a checkpoint, exact horizon, and input-length level are averaged in log response and actual context. Horizon-elasticity construction uses the corresponding mean log response. Four registered multi-capacity checkpoint subsets supply local capacity comparisons: Toto 2.0 (all five sizes), Moirai 1 (small, base, large), Chronos 2 (small, base), and

Chronos-Bolt (small, base). Within each subset, an ordinary least-squares slope of $\log R$ against $\log N$ estimates capacity elasticity at each common horizon–input-length condition containing at least two checkpoints. Indexing capacity targets by i, slopes at an input-length level are pooled across the available subsets and horizons; their median becomes the capacity target u_i, paired with their mean actual context L_i.

For each checkpoint and horizon, consecutive observed input-length levels produce changes in log response. The actual endpoint contexts and log-response changes are averaged over available horizons separately for each checkpoint and input-length transition. Indexing context targets by j, this gives log-response-change targets v_j at active capacity N_j and endpoints L_j^-, L_j^+. Let K and J count capacity and context targets, respectively. The complete fit uses $K = 8$ and $J = 115$; the leave-Toto-out fit uses $K = 8$ and $J = 90$. The predicted capacity and context targets, $\widehat{u}_i$ and $\widehat{v}_j$, are

$$\widehat{u}_i = -\beta \log L_i, \tag{15}$$

$$\widehat{v}_j = -(\gamma + \beta \log N_j) \log(L_j^+ / L_j^-) + q\big[(\log L_j^+)^2 - (\log L_j^-)^2\big]. \tag{16}$$

B.6 Parameter estimation and prediction metrics

The residual vector concatenates capacity residuals $r_i^{(N)}$ and context residuals $r_j^{(L)}$:

$$r_i^{(N)} = \frac{2(\widehat{u}_i - u_i)}{\sqrt{K}}, \qquad r_j^{(L)} = \frac{\widehat{v}_j - v_j}{\sqrt{J}}. \tag{17}$$

The factor two scales capacity residuals before robust loss evaluation. Let a index the concatenated residuals r_a, let $s = 0.02$ be the robust-loss scale, and define the soft-ℓ_1 loss $\rho(z) = 2(\sqrt{1+z} - 1)$ for a squared standardized residual z. Optimization uses the objective

$$\min_{(\gamma,\beta,q)\in\mathbb{R}^3} \ \frac{s^2}{2} \sum_a \rho\big((r_a/s)^2\big)\,. \tag{18}$$

Within each checkpoint and input-length level in the Controlled Grid-Horizon Panel, index the sorted scored horizons by k and define adjacent elasticities a_k:

$$a_k = \frac{\log R(H_{k+1}) - \log R(H_k)}{\log H_{k+1} - \log H_k}. \tag{19}$$

Hats denote parameter estimates. The estimate $\widehat{\alpha}$ is the median of these elasticities. For final combined Controlled points indexed by p, write (N_p, H_p, L_p, R_p) for each point's resource coordinates and observed response. The intercept estimate is then

$$\widehat{c} = \mathrm{median}_p \left[\log R_p - \widehat{\alpha} \log H_p + \widehat{\gamma} \log L_p + \widehat{\beta} \log N_p \log L_p - \widehat{q} (\log L_p)^2\right]. \tag{20}$$

The leave-Toto-out procedure removes every Toto 2.0 observation from both the response and horizon-evidence panels before reconstructing the local targets and re-estimating all five coefficients. In the dataset-held-out comparisons, response aggregation and local-target construction are repeated within each training partition, with validation responses aggregated separately (Appendix C.5).

For P response points, let $y_p = \log R_p$ and $\widehat{y}_p = \log \widehat{R}_p$ denote observed and predicted log responses. The absolute percentage error (APE) and its mean MAPE are

$$\mathrm{APE}_p = 100\,|\exp(\widehat{y}_p - y_p) - 1|\,, \tag{21}$$

$$\mathrm{MAPE} = P^{-1} \sum_p \mathrm{APE}_p. \tag{22}$$

Median APE is the empirical median of the pointwise errors. The Toto capacity-curve evaluation additionally averages observed and predicted responses over horizons at each fixed capacity and context, then computes APE on the five resulting capacity means.

C Additional experimental results of the Unified Scaling Law

C.1 Reference predictions and estimation stability

Table 5 compares the Unified law with a constant prediction, set to the exponential of the training-side median log response, and a capacity-independent form obtained by fixing $\beta = 0$ and re-estimating the remaining coefficients through the same procedure. All forms use the same evaluation points and exclude every Toto checkpoint when fitting their Toto predictions.

Table 5: Reference prediction MAPEs (%) on Controlled, Public, and held-out Toto response points.

Form	Controlled	Public	Toto
Constant	14.83	16.92	17.10
Capacity-independent	9.96	13.01	11.84
Unified	10.13	12.82	11.53

For coefficient and prediction stability, we resample the 15 base datasets with replacement 1,000 times, retaining all frequencies and observations of each draw, and repeat aggregation and every estimation step with the checkpoint panel fixed. The 95% percentile intervals are $[0.00090, 0.00322]$ for β and $[0.0131, 0.0310]$ for q, supporting history-dependent capacity gains and diminishing context returns.

Prediction comparisons use paired resamples for the Unified and capacity-independent forms, excluding all Toto checkpoints in each Toto fit. Evaluated on the fixed original response points, the 95% bootstrap ranges for Unified MAPE are $[9.59, 18.90]\%$ for Controlled, $[12.45, 21.05]\%$ for Public, and $[10.84, 20.11]\%$ for Toto. Paired MAPE-difference intervals include zero.

C.2 Complete Toto held-out evaluation

Figure 7 extends the main-text Toto evaluation to all 150 responses in the Controlled Grid-Horizon Panel, showing pointwise predictions and capacity profiles for the all-checkpoint and leave-Toto-out fits.

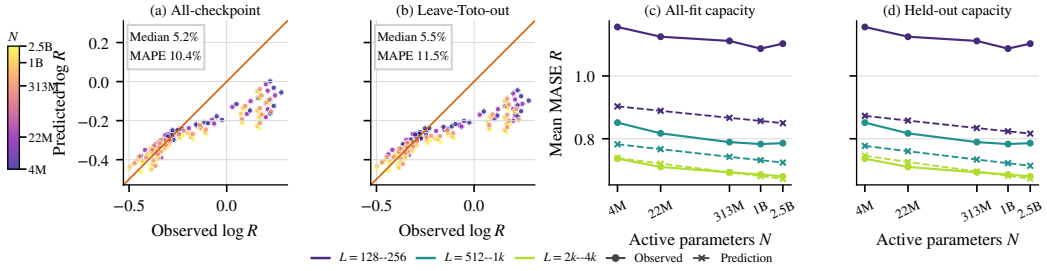

Figure 7: Complete Toto 2.0 evaluation. (a,b) Observed versus predicted $\log R$ at all 150 original NHL coordinates for the all-checkpoint and leave-Toto-out fits; color encodes active parameters and the diagonal denotes exact prediction. (c,d) Capacity response after pairing adjacent input lengths and averaging the five horizons. Solid circles are observed Toto means and dashed crosses are predictions from the corresponding fit. Each context curve contains all five Toto checkpoints.

C.3 Context-gain calculation and complete-support validation

For each fixed checkpoint and horizon, let R_{left} and R_{right} denote responses at shorter and longer inputs sharing the baseline denominator. The paired context gain in Figure 5 is $100(R_{\text{left}} - R_{\text{right}})/R_{\text{left}}$; positive values indicate improvement. The fitted curves illustrate the broad saturation region, with the 100M fit varying by about 1% across 4k–8k.

Using history and interface metadata, we select eight checkpoints and 12 tasks with complete recorded input support across all eight input lengths and five shared-origin Grid horizons, fixing task membership, forecast origins, and checkpoint membership. On this support set, mean paired gains are 3.46% for 1024–2048 and 1.33% for 2048–4096, becoming slightly negative for the two extensions beyond 4096. Repeating the estimation procedure gives saturation scales of approximately 3.6k–4.6k over 10M–1B parameters.

C.4 Qualitative restrictions of reduced laws

Figure 2 applies the restrictions in Section 5.2 to the full-panel law without refitting. Panel (a) uses adjacent-capacity checkpoints from the registered subsets at common horizons and input-length level $L = 1024$. Panel (b) subtracts the observed $L = 1024$ response separately within each checkpoint–horizon slice for the three checkpoints between 75M and 150M parameters. Panel (c) converts the local capacity slopes pooled across heterogeneous checkpoint subsets to tenfold log-response changes. Panel (d) divides adjacent-context log-response changes by their actual log-input change $\Delta \log L$ for checkpoints with $N \leq 50$M, whose median log capacity defines the representative 22M curve; the finite difference equals the quadratic law's derivative at the geometric midpoint.

C.5 Flexible alternatives and common estimation

We compare the Unified law with four alternatives that can express changing context returns and capacity–history coupling: the Unified law with an independent-capacity term (U+N), conditional quadratic (CQ), rational saturation (RS), and capacity-gated saturation (CGS), with log responses $f_{\mathrm{U+N}}$, f_{CQ}, f_{RS}, and f_{CGS}, respectively. Within these forms, $s > 0$ is a context scale in CQ and CGS; u and v control capacity and horizon dependence, respectively. In RS, $L_s > 0$ is the context saturation scale; in CGS, $N_0 > 0$ is the capacity-gate scale. The forms are

$$f_{\mathrm{U+N}} = c + \alpha h + \lambda n - \gamma \ell - \beta n \ell + q\ell^2, \tag{23}$$

$$f_{\mathrm{CQ}} = c + \lambda n + \alpha h + q(\ell - \log s - un - vh)^2, \tag{24}$$

$$f_{\mathrm{RS}} = c + \alpha h - (\gamma + \beta n)\frac{L}{L + L_s}, \tag{25}$$

$$f_{\mathrm{CGS}} = c + \alpha h - \gamma \frac{N}{N + N_0}\left[1 - \exp\left(-\frac{L}{sH^v}\right)\right]. \tag{26}$$

The first is the six-parameter nested extension. CQ uses a capacity- and horizon-dependent quadratic center and has seven parameters. RS has five parameters and an explicit context scale. CGS has six parameters and combines a Hill capacity gate with a horizon-dependent saturation scale. Parameter counts include c and α.

The comparison uses the 23-task Controlled Grid-Horizon Panel and its 16 development checkpoints; the complete Toto 2.0 checkpoint series is excluded. The 15 base datasets form three fixed folds, keeping every frequency and checkpoint–horizon–context observation of a base dataset in the same fold. Training and validation responses are aggregated separately. Each candidate fits its implied local relations using the same targets, weighting, and robust objective; the intercept is calibrated last. Optimization and initialization use training data only.

Mean validation MAPEs across the three folds are 11.77% for Unified, 11.78% for its independent-capacity extension, 11.87% for CQ, 11.90% for RS, and 13.10% for CGS. Fold assignments and per-fold results are provided in the released analysis.

C.6 Horizon-interaction extensions

Let f_{U} denote the Unified log-response law in Eq. 5, and let η and κ be capacity–horizon and context–horizon interaction coefficients. We assess the common-horizon-shift specification with the nested family

$$f = f_{\mathrm{U}} + \eta n h + \kappa \ell h. \tag{27}$$

We allow η alone, κ alone, or both. These extensions have six, six, and seven parameters. The nh term permits horizon-dependent capacity gains, while ℓh permits a horizon-dependent context saturation scale.

All forms use the same dataset folds and jointly fit matched capacity, context, and horizon contrasts. Validation errors remain comparable across these forms. The released code and results provide the fitting objective and fold-level comparisons.

D Unified Theory of Time Series Learning: theoretical analysis

D.1 Gaussian learning model

We analyze a fixed query B and scalar forecast target Y. Conditional on this query, let

$$Y = \sum_{j=1}^{p} \psi_j(B)\omega_j + \varepsilon, \qquad \omega_j \sim \mathcal{N}(0, \tau_j^2), \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2). \tag{28}$$

The query features $\psi_j(B)$ are fixed, and the regression coefficients and innovation are independent. Let $z_{L,j}$ denote the sufficient statistic supplied by historical observations for coefficient j, with independent standard Gaussian noises ξ_j also independent of the coefficients and innovation:

$$z_{L,j} = \omega_j + \frac{\xi_j}{\sqrt{L\nu_j}}, \qquad \xi_j \sim \mathcal{N}(0,1), \quad \nu_j > 0, \tag{29}$$

Thus $L\nu_j$ is the precision of the historical evidence about coefficient j. Gaussian conditioning gives the posterior mean $\widehat{\omega}_j(L)$ and variance $v_j(L)$ (Rasmussen & Williams, 2006):

$$\widehat{\omega}_j(L) = \frac{L\nu_j\tau_j^2}{1 + L\nu_j\tau_j^2} z_{L,j}, \qquad v_j(L) = \frac{\tau_j^2}{1 + L\nu_j\tau_j^2}. \tag{30}$$

Consider nested predictors that use the posterior means of the first r coefficients and the prior means of the remainder. Here r counts the available regression coefficients. The coefficients ω_j describe the dynamic rule; the fitted or pretrained neural parameters determine how an actual model estimates and applies such a rule. Under the joint distribution of coefficients, evidence, and innovation at fixed B, the prediction error is a centered Gaussian with variance

$$V_r(L) = \sigma^2 + \sum_{j \le r} \psi_j(B)^2 v_j(L) + \sum_{j > r} \psi_j(B)^2 \tau_j^2. \tag{31}$$

The posterior mean is also the posterior median, so the minimum expected absolute error among predictors using these r statistics is

$$\mathcal{E}_r(L) = \sqrt{\frac{2}{\pi}} \sqrt{V_r(L)}. \tag{32}$$

The estimate can be retained after training or computed from current history, as in in-context regression of autoregressive rules (Sander et al., 2024).

D.2 Historical evidence and available rule coefficients

Write $a_j = \psi_j(B)^2\tau_j^2$ and $b_j = \nu_j\tau_j^2$. Coefficients with $a_j = 0$ do not affect the fixed query and can be omitted. Define the remaining variance $A_r = \sigma^2 + \sum_{j>r} a_j$. Then

$$V_r(L) = A_r + \sum_{j \le r} \frac{a_j}{1 + b_j L}. \tag{33}$$

We assume $A_r > 0$ for the predictors under comparison. This remaining variance includes innovation and the contribution of coefficients outside the predictor's available set. Denote the variance reduction from adding coefficient $j = r + 1$ by $g_j(L)$:

$$g_j(L) = V_r(L) - V_{r+1}(L) = \frac{a_j b_j L}{1 + b_j L}. \tag{34}$$

Its absolute-error improvement is

$$\mathcal{E}_r(L) - \mathcal{E}_{r+1}(L) = \sqrt{\frac{2}{\pi}} \frac{g_j(L)}{\sqrt{V_r(L)} + \sqrt{V_{r+1}(L)}}. \tag{35}$$

The numerator increases with L, while both variances in the denominator are nonincreasing. The improvement therefore increases with historical evidence whenever the added coefficient affects the query.

For fixed r and finite p, put $B_r = \sum_{j \le r} a_j/b_j$. Writing $O(L^{-2})$ for a remainder bounded in magnitude by a constant times L^{-2} at large L, expanding Eq. 33 gives

$$V_r(L) = A_r + \frac{B_r}{L} + O(L^{-2}), \qquad \mathcal{E}_r(L) = \sqrt{\frac{2A_r}{\pi}} \left(1 + \frac{B_r}{2A_r L} + O(L^{-2})\right). \tag{36}$$

The finite-history estimation error decreases while the remaining variance dominates. In particular, the magnitude of the log-context derivative of error approaches zero.

D.3 Retained rule information and immediate evidence

For one coefficient, suppose training leaves a Gaussian posterior with variance $v > 0$, and immediate historical evidence contributes an additional Gaussian likelihood with conditional precision $I \ge 0$. Gaussian conditioning gives the updated posterior variance v_{post}:

$$v_{\mathrm{post}} = \frac{v}{1 + Iv}. \tag{37}$$

Here I measures information additional to the retained estimate, accounting for overlap between the training record and current history by conditioning on retained evidence. For fixed query weight $k > 0$ and remaining variance $A \ge 0$, the absolute-error gain $G(v; I)$ from that immediate evidence is

$$G(v; I) = \sqrt{\frac{2}{\pi}} \left[\sqrt{A + kv} - \sqrt{A + \frac{kv}{1 + Iv}}\right]. \tag{38}$$

With I, k, and A fixed, this gain is nondecreasing in v. Indeed,

$$\frac{\partial G}{\partial v} = \frac{k}{\sqrt{2\pi}} \left[\frac{1}{\sqrt{A + kv}} - \frac{1}{(1 + Iv)^2 \sqrt{A + kv/(1 + Iv)}}\right] \ge 0, \tag{39}$$

because $(1 + Iv)^4[A + kv/(1 + Iv)] \ge A + kv$, with strict inequality for $I > 0$. More accurate retention of the same coefficient therefore reduces the gain from additional immediate evidence. Comparing two positive precisions follows by first incorporating the smaller precision into v and applying the same argument to the additional precision.

E Unified Theory of Time Series Learning: experimental evidence

E.1 History budgets, immediate input, and adaptation

The matched-history comparison uses DLinear and PatchTST adapted from Time-Series-Library (Wu et al., 2023). A shared univariate model is trained on each task's scalar histories, using the available length-L record before the earliest scoring origin, 96-step inputs, and H-step targets. Adam uses learning rate 10^{-4}, batch size 32, and three initialization seeds. Chronological validation selects an epoch count, followed by refitting on the complete pre-origin record; weights remain fixed during scoring. The comparison contains 1,656 full-shot cells. Each plotted response averages the 23 tasks and then the seeds equally; seed variability is the sample standard deviation relative to the seed mean at each horizon–history condition. TSFM curves summarize the same nine checkpoints covering the complete grid, each represented by its 23-task arithmetic mean. Remaining architecture, normalization, and optimization settings are specified in the released configuration.

The crossed experiment varies training history $T \in \{4096, 8192\}$ and immediate input $C \in \{96, 512\}$ at $H = 96$ on the same 23 tasks. DLinear, PatchTST, a compact PatchTST, and pretrained TimesFM 2.5 each use three seeds, giving 1104 fits. Histories end before the earliest scoring origin. Training starts and validation targets are matched across input lengths by cropping the corresponding 512-step windows for shorter inputs; evaluation origins and targets are identical across arms. Each input configuration is trained separately under a common update budget, with validation

Table 6: Validation-selected MASE in the crossed experiment, averaged over 23 tasks and three seeds. T is training history and C is immediate input.

Model	Input C	History $T = 4096$	History $T = 8192$
DLinear	96	1.04264	1.03958
DLinear	512	0.88869	0.88328
PatchTST	96	0.91400	0.91361
PatchTST	512	0.86638	0.84896
PatchTST-compact	96	0.99099	0.98986
PatchTST-compact	512	0.90357	0.89159
TimesFM	96	0.83326	0.81896
TimesFM	512	0.74498	0.75370

selecting among the initial and subsequent predictors without refitting. MASE uses the same GIFT scoring and matched Seasonal Naive reference as the main evaluation. Tasks are averaged equally, with seeds and the balanced history–input cells averaged within task.

Adapted TimesFM improves MASE on 20 tasks, is unchanged on two, and increases it on one. Averaging training histories and seeds within task, frozen TimesFM has MASE 1.0431 and 0.8035 at inputs 96 and 512; the adapted configurations have MASE 0.8261 and 0.7493. The short–long input gap decreases from 0.2396 to 0.0768. Its reduction of 0.1628 has a 95% bootstrap interval of $[0.0170, 0.3248]$, obtained from 10,000 resamples of the 15 base datasets while keeping their frequency tasks together.

E.2 Controlled processes, evidence splits, and scoring

Writing sign for the sign function, the controlled processes with rule sign $s \in \{-1, +1\}$, magnitude $a \in \{0.15, 0.225, 0.3\}$, and independent standard Gaussian innovations ϵ_t are

$$\text{AR(1):} \quad x_t = sa\, x_{t-1} + \sqrt{1 - a^2}\epsilon_t, \tag{40}$$

$$\text{Lag-8:} \quad x_t = sa\, x_{t-8} + \sqrt{1 - a^2}\epsilon_t, \tag{41}$$

$$\text{Threshold:} \quad x_t = sa\ \text{sign}(x_{t-1}) + \sqrt{1 - a^2}\epsilon_t. \tag{42}$$

AR(1) and lag-8 have standard Gaussian stationary marginals; lag-8 starts with eight independent standard Gaussian observations. The threshold process starts from its stationary equal mixture of $\mathcal{N}(-a, 1 - a^2)$ and $\mathcal{N}(a, 1 - a^2)$. Every process has mean zero and variance one, and the two signs have the same marginal distribution within each process and magnitude.

Independent 96-step queries are drawn from the equal mixture of the two stationary rules and retained when their absolute rule log-likelihood ratio is at most $\log 3$, selecting queries that leave uncertainty about the rule. Each query has an antithetic partner obtained by negating the entire series. Frozen-context and parameter experiments use 128 independent development queries per process at $a = 0.3$ and 1,024 independent test queries per process and magnitude. Historical prefixes are sampled from the exact reverse conditional process given the query boundary; frozen-context comparisons use nested lengths and four prefix realizations. Activation experiments use separate query streams, with 128 development and 1,024 confirmation queries per process and magnitude, and two donor repetitions. Development and confirmation donors are independent; the primary TimesFM setting and candidate grids are fixed before confirmation.

Let b_j, $j = 1, \ldots, 96$, denote the query observations. The primary target is the one-step conditional mean: $sa\, b_{96}$ for AR(1), $sa\, b_{89}$ for lag-8, and $sa\ \text{sign}(b_{96})$ for threshold. MASE follows Eq. 1, with both model and naive forecasts scored against this target. The raw absolute-error scale is the mean lag-one absolute difference within the common 96-step query; the naive forecast is b_{96}. Raw scores are pooled across paired rule signs before division by the naive score, then averaged across conditions. Paired comparisons retain antithetic queries in the same statistical unit and average prefix-level errors when prefixes are repeated. The 95% intervals use 2,000 bootstrap resamples of independent queries, preserving their paired predictions and histories. Training seeds are reported separately; query intervals condition on the fitted models and donor states. The decomposition in Eq. 46 uses the same predictions.

E.3 HISTORICAL EVIDENCE AND RULE IDENTIFICATION

Chronos 2 and TimesFM 2.5 are evaluated at context lengths 96, 128, 256, 512, 1024, and 2048. Each emits its native median forecast from the first output block, with no rollout or state shared across examples. We compare native normalization statistics with fixed population location zero and scale one; Chronos retains its nonlinear scaling transform. Interventions use fixed statistics throughout. Using subscripts $+$ and $-$ for rule signs $s = +1$ and $s = -1$, respectively, denote predictions by $f_+(B), f_-(B)$ and conditional means by $\mu_+(B), \mu_-(B)$. Summing over evaluated queries B, define the rule projection P:

$$P = \frac{\sum_B [f_+(B) - f_-(B)][\mu_+(B) - \mu_-(B)]}{\sum_B [\mu_+(B) - \mu_-(B)]^2}. \tag{43}$$

It measures response direction and amplitude, with one corresponding to the oracle contrast amplitude. Table 7 shows prediction gains and differences between models' rule responses. Across both normalization conditions, all 36 model–process–magnitude–normalization combinations have positive intervals for the MASE gain from 96 to 2048.

Table 7: Conditional-mean MASE with fixed population statistics at $a = 0.3$. Rule projection measures the predicted contrast along the known conditional-mean contrast; one corresponds to the oracle amplitude.

Process	Frozen model	$L = 96$	$L = 2048$	Rule projection
AR(1)	TimesFM 2.5	0.32424	0.12650	0.812
AR(1)	Chronos 2	0.33295	0.12958	0.666
Lag-8	TimesFM 2.5	0.33272	0.16004	0.739
Lag-8	Chronos 2	0.34423	0.26440	0.133
Threshold	TimesFM 2.5	0.37300	0.22949	0.454
Threshold	Chronos 2	0.37281	0.25215	0.369

An oracle knows the process family and coefficient magnitude and estimates the rule sign from the historical prefix, using the same queries, conditional histories, and scoring as the frozen models. Write D_L for the history $D_{L,t}$ containing query B, with the forecast-origin index suppressed, and let p_+ and p_- denote sequence densities under the positive and negative rules. Because both signs are equally weighted for each fixed query B, the log posterior odds $\Lambda(D_L \mid B)$ are

$$\Lambda(D_L \mid B) = \log \frac{p_+(D_L)}{p_-(D_L)} - \log \frac{p_+(B)}{p_-(B)}. \tag{44}$$

Selecting the more probable sign gives a posterior-median prediction and minimizes expected absolute error; ties select the positive sign. At the smallest magnitude, $a = 0.15$, its MASE is below 0.001 for all three processes at $L = 512$. All processes and magnitudes have zero sign errors at lengths 1,024 and 2,048 in the registered finite sample. The released results contain all 54 process–magnitude–length conditions.

E.4 PARAMETER RETENTION AND TRANSFER

DLinear and a compact PatchTST learn the controlled processes with fixed population normalization, 96 inputs, three outputs, and mean squared error (MSE) loss. Histories have lengths 1024, 4096, or 8192, with three training seeds and opposite-rule pairs, giving 108 individual fits. Histories are nested, and opposite rules share initialization and sampling randomness. An 80% chronological training split supplies windows; the remaining targets select the predictor without refitting. Architecture and optimizer settings are specified in the released configuration.

Trained models predict independent queries after the training record has left the forecast computation. We exchange parameters between opposite-rule models and calibrate their batch-normalization (BatchNorm) statistics, where present, on the same independent unlabelled mixture of historical windows for both directions, without future targets or parameter updates. Error to the donor rule decreases in all 54 process–architecture–length–seed conditions. Both architectures' mean prediction errors decrease with training-history length (Table 8).

Table 8: Conditional-mean MASE of trained predictors, averaged over three seeds. Transfer projection measures the donor-rule response after parameter exchange and shared-history normalization calibration at $L = 8192$.

Process	Model	$L = 1024$	$L = 4096$	$L = 8192$	Transfer projection
AR(1)	DLinear	0.27871	0.16235	0.12634	0.986
AR(1)	PatchTST	0.32585	0.14790	0.09567	1.028
Lag-8	DLinear	0.32463	0.17748	0.12204	0.985
Lag-8	PatchTST	0.34635	0.14133	0.10034	0.946
Threshold	DLinear	0.35156	0.25287	0.22841	0.584
Threshold	PatchTST	0.37115	0.20446	0.18716	0.755

E.5 Activation transfer, recovery, and controls

The primary TimesFM 2.5 intervention uses activation patching (Zhang & Nanda, 2024) at the sixth Transformer block's residual output, over the final three query tokens, with input length 512 and unit strength. These settings were fixed on development evidence before independent confirmation. Each recipient uses four independent 512-step donor histories from the same process and rule, with two donor repetitions. Chronos 2 uses the same donor construction and history lengths, patching the six tokens corresponding to the final 96 observations while leaving its register and forecast tokens unchanged. Its layer and strength are selected on development evidence as specified below.

For donor index $j \in \{1, \dots, 4\}$, let D_j denote donor history and B_j its final 96 observations. Let h map each input to the selected donor-token states. We construct the mean state difference Δh:

$$\Delta h = \frac{1}{4} \sum_{j=1}^{4} \big[h(D_j) - h(B_j) \big]. \tag{45}$$

Only observed donor histories enter this construction; known synthetic rules determine matched and opposite donor assignment and scoring targets. With intervention strength λ, transfer replaces a new query's short-input state by $h(B) + \lambda \Delta h$. Recovery first replaces the selected state in its natural 512-step execution by $h(B)$, defining erasure, and then adds $\lambda \Delta h$. Other long-context states, including prefixes accessible to downstream layers, remain unchanged.

Controls match donor count, strength, and location. Opposite-rule donors hold the recipient query and long history fixed and pair donor carriers and sampling randomness across signs. Random changes match the norm of Δh; carrier-only changes compare independent short donor queries; shuffled-history changes permute each donor's earlier observations while preserving its final 96 values. Natural predictions and norm-matched random erasure are also retained.

TimesFM is evaluated at blocks 4, 6, and 8 with strengths $0.5, 1, 1.5$, retaining all nine settings. Chronos development evaluates blocks 3, 6, and 9 at the same strengths. Equal averaging of development conditional-mean MASE across processes, magnitudes, repetitions, signs, and both objectives selects block 6 and strength 1.5, which are then fixed for independent confirmation. Each setting contains 18 process–magnitude–donor-repetition conditions sharing the registered query sets.

Table 9: Conditional-mean MASE for the primary TimesFM setting and independently confirmed Chronos setting. Baseline is query-only prediction for transfer and erased prediction for recovery. Matched and opposite refer to the donor rule.

Model	Objective	Baseline	Matched	Opposite	Random	Shuffled
TimesFM 2.5	Transfer	0.27810	0.25720	0.30316	0.31468	0.26804
TimesFM 2.5	Recovery	0.25821	0.23726	0.29228	0.29514	0.25137
Chronos 2	Transfer	0.28774	0.28776	0.29987	0.28710	0.29144
Chronos 2	Recovery	0.21064	0.19798	0.20436	0.20984	0.20454

Matching donors outperform opposite-rule donors in all 324 TimesFM comparisons across nine settings, two objectives, and 18 conditions, each with a positive paired-query interval. Their prediction changes also align with the oracle rule contrast in every condition. Mean recovery improves at all

nine settings and mean direct transfer at seven; the released code and results contain the complete grid. At the primary setting, MASE is lower than with shuffled-history donors in 14 of 18 conditions for each objective. The recovery percentage in Section 6 is mean recovered improvement divided by mean erasure loss at that setting.

Chronos recovery improves all 18 conditions, each with a positive paired interval. Matching donors outperform opposite-rule donors in all 18 conditions for each objective, with positive intervals in 15. Its direct-transfer MASE is nearly unchanged, from 0.28774 to 0.28776 (Table 9).

To separate rule responses from common prediction changes, let the two conditional means at horizon 1 be $\pm q(B)$ and predictions under the corresponding history or donor rules be $f_+(B), f_-(B)$. Suppressing the fixed query argument B, define the common prediction component $m = (f_+ + f_-)/2$ and the rule-contrast component $d = (f_+ - f_-)/2$. The paired-rule squared error decomposes exactly as

$$\frac{(f_+ - q)^2 + (f_- + q)^2}{2} = m^2 + (d - q)^2. \tag{46}$$

The first term measures a common prediction offset and the second agreement with the rule-dependent conditional mean. Gains are relative to query-only prediction for transfer and erasure for recovery. TimesFM's rule-dependent error improves in every condition for both objectives at the primary setting, including against shuffled-history donors, with positive paired intervals. Chronos improves the mean rule component in each process for both objectives. In direct transfer, its common-component cost exceeds this gain in MSE; recovery lowers overall error.