ZERO-SHOT GENERALIZATION WITH COVLIFT: CONTEXT AND COVARIATE CAPACITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models (TSFMs) promise accurate zero-shot forecasts on unseen series. Frozen models can nevertheless fail to use predictive information already available in their inputs. We investigate how context length shapes the value of covariates and develop **CovLift**, a method for adapting inputs to frozen TSFMs. CovLift screens engineered covariates against observed history and matches features and context length to the forecasting model. Synthetic probes predict Chronos-2's covariate-to-numeric-only error ratio on unseen relationships. On spring electricity load with day-1 weather forecasts, eight added features harm Chronos-2 at 128 observations but reduce its geometric paired mean squared error ratio by 24% against same-length raw covariates at $1{,}024$. This context effect replicates across sixteen unseen regions, persists under an assumed six-hour publication delay, and has the same direction on TiRex-2. CovLift also improves over raw covariates on unseen load regions. Summer evidence is inconclusive; tested additions show no confirmed Chronos-2 gain on solar or wind, and the original CovLift-Auto configuration slightly worsens aggregate benchmark performance. These findings identify context length as a determinant of covariate value and demonstrate how input design can improve zero-shot generalization on unseen load regions without updating model weights.

1 INTRODUCTION

TSFMs (Ansari et al., 2024; Woo et al., 2024; Das et al., 2024; Ansari et al., 2025) forecast new series *zero-shot*, without fine-tuning or adapters. Fine-tuning can underperform training from scratch (Zhang et al., 2026). We ask how a frozen *reader* uses predictive inputs. A day-ahead temperature forecast cuts error tenfold for a temperature target (§4), yet can harm other targets. We study this *under-generalization* across synthetic probes, energy, retail and benchmark tasks (§3–4; Fig. 1).

Longer context turns engineered covariates from harmful to useful on spring load. Across fourteen Rounds-2/3 regions, 328 of 338 windows qualify at every L. Top-eight expansion on the day-1 weather feed harms Chronos-2 at $L{=}128$ (1.12) but reduces its geometric paired MSE ratio by 24% at $L{=}1{,}024$, against same-length raw covariates (preregistered contrast 0.388 $[0.290, 0.494]$). The effect replicates under an assumed six-hour publication delay on the corrected feed and on 16 unseen regions; TiRex-2 follows at a quarter of the size (Fig. 3; §6).

Holding feature engineering fixed exposes reader differences. CovLift uses a standard feature dictionary and correlation screen. A generic control with the same dictionary and screen selects identical reader inputs at all 543 audited Chronos-2 origins, and ten registered selection studies did not establish a deployable rule that beats the screen (App. D.1). The screen measures explained target variation; probes and context sweeps test how readers use it.

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.

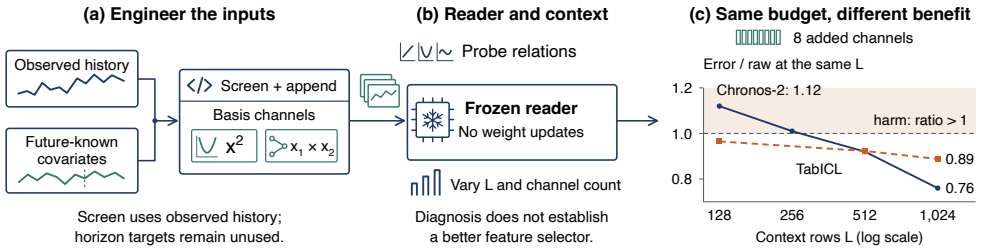

Figure 1: **Feature value depends on reader and context.** Eight appended channels change Chronos-2's forecasts from harm to benefit. Curves: geometric paired MSE ratios against same-length raw covariates, 328 spring load windows, legacy day-1 feed; lower is better. The channel budget is fixed; channel identities are reselected at each L. Intervals and fixed-identity controls: Fig. 3.

Contributions.

- **We establish an empirical regularity linking context length to feature value.** On spring load, Chronos-2's deployed screen moves from 0.96 to 0.73 of same-length raw-covariate error between L=128 and 1,024 (geometric MSE ratios). The context effect survives fixed channel identities and replicates on 16 unseen regions and in sign on TiRex-2 (§6, Fig. 3).
- **We validate profile-based predictions of Chronos-2's covariate use.** Probe predictions fixed before testing rank Chronos-2's covariate-to-numeric-only error ratio on eight held-out links (ρ=0.952). Within-window correlation predicts them comparably well; the profile predictor fails on TabICL (§4).
- **We develop CovLift, a training-free intervention for frozen forecasters.** CovLift matches the feature dictionary and context length to the reader. Its screened weather-only expansion at L=1,024 beats the deployed configuration on unseen load regions (preregistered geometric MSE ratio 0.841 $[0.794, 0.891]$). Solar and wind mark a boundary: no tested channel set yields a confirmed gain for Chronos-2 (§5, §6).

2 Related Work

Covariate readers. TSFMs use adapters (ChronosX (Pineda Arango et al., 2025), UniCA (Han et al., 2026), CoRA (Cheng et al., 2026)) or native readers (Chronos-2 (Ansari et al., 2025), TiRex-2 (Podest et al., 2026)). TabPFN-TS (Hoo et al., 2026) combines temporal features with pretrained TabPFN-v2; TimeXer (Wang et al., 2024a) is trained per dataset. Berthelier et al. (2026), closest in aim, probe Chronos-2 and TabPFN-TS on fixed hand-built relations and report covariate harm. We predict the FM-own error ratio on eight held-out links, with predictions fixed before testing, and measure feature value by context length (§4, §6).

Test-time methods. With frozen TSFM weights, methods fit a per-instance GP cascade (TFMAdapter (Dange & Sarawagi, 2025)), optimize *target-history* transforms (TATO (Qiu et al., 2026)), or calibrate at test time (Chen & Liang, 2025). Prop. A2 characterizes affine-equivariant target-history pipelines. Learned methods address drift, representation alignment, residual correction, or evolving domains (Zhao et al., 2026; Hu et al., 2026b; Liang et al., 2026; Zeng et al., 2024). TAFAS and COSA update adapters on newly observed targets (Kim et al., 2025; Im & Kwon, 2026).

Input engineering. TS-RAG retrieves time-series patterns (Ning et al., 2025). Crafter mines the frozen residual for corrective features admitted on a validation score (Wang et al., 2026a), GITCO gates context patches with a classifier trained across datasets (Pandey et al., 2026), and exogenous dropout is a training-time covariate-robustness intervention (Hu & Ai, 2026).

Feature engineering. CovLift uses basis features (Oreshkin et al., 2020) and residual-correlation screening. Generate-then-evaluate pipelines include CAAFE (Hollmann et al., 2023) and OpenFE (Zhang et al., 2023). We test feature value as context observations per channel increase (§6).

Frozen normalization. RevIN (Kim et al., 2022) and later normalizers are evaluated with trained forecasters (Fan et al., 2023; Liu et al., 2023; 2022; Han et al., 2024; Yang et al., 2026; Lu et al., 2024;

Hu et al., 2026a). We test nine frozen backbones and toggle a tenth model's normalizer with weights fixed, measuring which shifts normalization absorbs (cf. Toner & Darlow, 2024) (§3; App. B.3).

Evaluation controls. Contamination detection (Shi et al., 2024; Oren et al., 2024) and post-cutoff evaluation (Golchin & Surdeanu, 2024; White et al., 2025; Roberts et al., 2024) inform our controls. We follow TS-Arena's preregistration practice (Meyer et al., 2026) and account for autocorrelated horizons (Wang et al., 2026b). Leakage catalogues address contamination (Meyer et al., 2025); vintage-consistent pretraining addresses revision bias (Carriero et al., 2026). We specify covariate availability at the forecast origin; the OpenSTEF Liander 2024 card records forecast availability time (OpenSTEF, 2025).

In-context limits. In-context learners favor function classes seen in pretraining (Wang et al., 2025; Bhattamishra et al., 2024); synthetic tasks establish sample-size dependence of in-context regression (Wang et al., 2024b; Akyürek et al., 2023; Bordelon et al., 2026). Adaptation analyses address non-stationarity (Qin et al., 2026). We test production readers through held-out-link prediction and context sweeps. The log-ratio context contrast is $7.5\times$ a stylized least-squares variance reference. It is not a matched empirical least-squares result.

3 Setup: Forecasting Interfaces and Metrics

Frozen forecasting interfaces. A rolling window contains target history $\mathbf{x} \in \mathbb{R}^L$ and future targets $\mathbf{y} \in \mathbb{R}^H$: L is context length and H the forecast horizon (L=512, H=96 unless stated). The *numeric-only* interface $f_{\text{num}}(\mathbf{x})$ reads target history. The *covariate* interface $f_{\text{cov}}(\mathbf{x}, \mathbf{C})$ also reads $\mathbf{C} \in \mathbb{R}^{(L+H)\times K_0}$, with K_0 native covariate channels; §6 specifies their forecast vintages. Both use the same frozen checkpoint. *Raw* covariates are the supplied channels; *compiled* covariates are CovLift's transformed inputs.

Metrics and comparators. For forecast $\hat{\mathbf{y}} = f(\cdot)$, $\text{MSE}(f) = H^{-1}\sum_{h=1}^{H}(y_h - \hat{y}_h)^2$ and skill $= \text{MSE}(\text{model})/(\min_{b\in\mathcal{B}} \text{MSE}(b) + \epsilon)$. The battery specifies $\mathcal{B}$ and ϵ (App. B); baselines include naive, drift (Hyndman & Athanasopoulos, 2021), damped drift (Gardner & McKenzie, 1985) and seasonal-naive. The scoring baseline is selected retrospectively; tables identify pooled-score exceptions.

Metric	Comparison	Aggregation	Favorable
Skill	Model / best baseline	Median of per-window skill	Lower
FM-own ratio	Covariate / numeric-only; same reader	Ratio of median skills: $t_{\text{cov}}/t_{\text{num}}$	Lower
Geometric MSE ratio	Usually compiled / raw	Exponentiated mean paired log-MSE difference	Lower
Surplus	Fitted affine / reader	Median paired log-MSE ratio	Positive

For the FM-own ratio, $t_{\text{num}} = \text{median}_w \, \text{skill}(f_{\text{num}})$ over windows w; t_{cov} is defined likewise. Shared per-window baselines generally do not cancel; intervals use moving-block bootstrap (Künsch, 1989). Geometric ratios retain the scorer's numerical guards. Captions specify comparators and weighting: windows for context sweeps, equal-weighted series for prospective rounds (§6). Positive surplus favors the reader (§4, App. B).

Controls for profile gains. A lower FM-own ratio alone does not establish new information or exclude memorization. A *shuffle control* deranges horizon covariates across windows and must raise $t_{\text{cov}}/t_{\text{num}}$ by >0.05. Chronos-2 profile gains must also replicate on the decontaminated `chronos-2-synth` checkpoint. CovLift evaluations use the separate registered audits in §6.

Normalization preserves structure, not accuracy. Exact mean/std normalization (RevIN (Kim et al., 2022)) maps a nonconstant context to $\mathbf{z} = (\mathbf{x} - \mu(\mathbf{x}))/\sigma(\mathbf{x})$ and reverses this operation after forecasting. It removes positive affine level and scale changes exactly; implementations differ (Prop. A1, App. B). For $x_t = a + bt + \eta_t$, the normalized window retains residual shape and trend coefficient $\tau = b\, s_L/\sigma$, where s_L is the standard deviation of the context time indices. Noiseless trends of the same sign become one ramp that an equivariant pipeline can extrapolate.

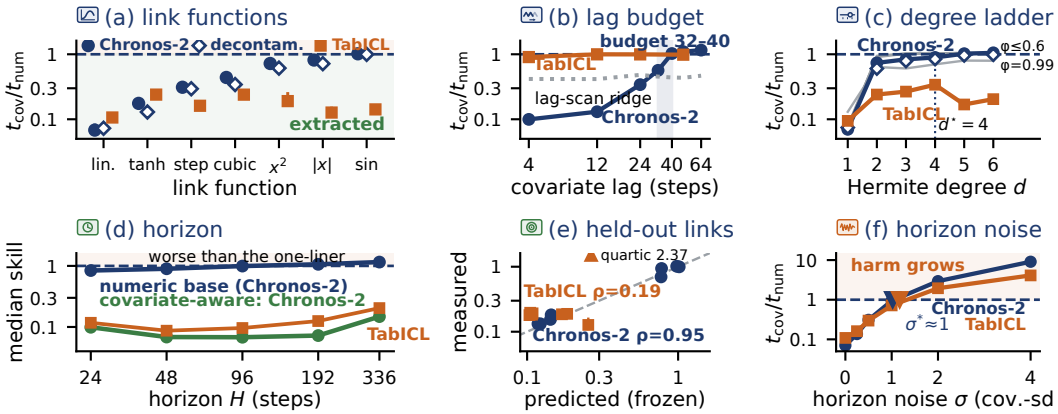

Figure 2: **The capacity profile** ($t_{\text{cov}}/t_{\text{num}}$; < 1 = the covariate helps beyond the model's own numeric forecast; whiskers: 95% moving-block CIs). **(a)** Links: Chronos-2's performance declines from linear to sin; TabICL reads all seven calibration links. **(b)** Lag: Chronos-2's performance declines between 24 and 40 steps; TabICL shows only AR carry-over. **(c)** Hermite degree at φ=0.9: Chronos-2's performance declines monotonically to $d^\star$=4; TabICL shows no cutoff; thin lines show the smoothness sweep. **(d)** Horizon (linear link): numeric-only performance declines with H; Chronos-2's relative gain shows no clear trend. **(e)** Held-out links (n=8 per reader), predictions frozen in advance: they track Chronos-2 (ρ=0.952) and fail on TabICL (ρ=0.19), where the quartic link harms the reader (2.37, clipped). **(f)** Horizon noise: break-even $\sigma^\star$ is near one covariate standard deviation; harm grows beyond it. Exact values: Table 3.

Any de-normalized readout using only the normalized context and activations computed from it is affine-equivariant (class $\mathcal{I}$ in App. B). The **affine-orbit factorization** (Prop. A2) writes it as $F(\mathbf{x}) = \sigma(\mathbf{x})\, G(N(\mathbf{x})) + \mu(\mathbf{x})$, where N is standardization. The factorization describes invariance; forecast accuracy depends on the learned mapping G. Future-known covariates can add information while preserving target equivariance (Prop. A2(b)).

Across nine frozen backbones, six graded shift types, eight datasets and five seeds, we observe level/scale robustness and trend degradation (Table 2). Removing normalization raises Timer's (Liu et al., 2024) level-shift error ratio from 1.00 to $\mathbf{83.9}\times$.

4 Measuring What the Covariate Reader Can Use

We vary one covariate factor at a time (n=400 windows per baseline cell, L=512, H=96), comparing Chronos-2's group-attention path over future-known channels (Ansari et al., 2025) with TabICLv2's row-wise in-context regressor (Qu et al., 2026; 2025). The *capacity profile* summarizes the responses (Fig. 2); lower $t_{\text{cov}}/t_{\text{num}}$ means greater benefit relative to numeric-only forecasting. Shuffle and decontamination controls follow §3; we also test two real datasets. For this checkpoint and sampling regime, the hash-locked profile sets CovLift's transform families, native lag budget, horizon-noise break-even $\sigma^\star$ and channel cap (§5), which changes with context length (§6).

Response to link shape. Chronos-2 reads *monotone* links, performs worse on symmetric non-monotone ones (quad, abs), and fails on sin and e_1e_2 (Fig. 2a), although covariates encode these relations. The seven-link ordering survives SNR 0.5 and decontamination (App. C); TabICL reads all seven calibration links without a clear form ordering.

Response to degree and smoothness. Orthogonal probabilist–Hermite probes He_d isolate degree d. At φ=0.9, Chronos-2 degrades *monotonically* with pure degree, reaching $d^\star$=4 (Fig. 2c). The cubic ratio (0.440) suggests reliance on its degree-1 projection (He_3 alone: 0.836). A registered φ sweep rejects the polynomial-kernel explanation; the preregistered window-geometry alternative remains consistent: $d^\star$ depends on smoothness (Fig. 2c, Table 5).

Prediction on held-out links. We test predictions on new links within this regime (Zhou et al., 2026; Ruan et al., 2024). A link's Hermite spectrum describes its degree composition (the *demand profile*); the measured curve describes the reader's response (the *ability profile*). Their pairing predicts $t_{\mathrm{cov}}/t_{\mathrm{num}}$, the reader's covariate-to-numeric ratio of median baseline-normalized errors, on held-out links. For Chronos-2 with raw covariates, predictions frozen before testing achieved Spearman 1.000 in sample and ρ=0.952 on eight *held-out* links (Fig. 2e; p=5.7×10^{-4}; Table 4).

For this FM-own ratio, the profile prediction has MAE 0.046 with no fitted parameter. The fitted comparators reach 0.136 (effective degree) and 0.048 (raw correlation). Six mixture cells expose the affine map's limits: it holds within 0.033 on He_2 mixtures but misses He_3 mixtures by 0.10–0.40. We therefore claim no functional form. TiRex-2 shares the ordering (Spearman -0.911 with within-window correlation) and reads the even links Chronos-2 mostly cannot (quad 0.275 against 0.722).

Comparison with a fitted affine reference. Our *linear-reader reference* fits an affine forecast within each window; the recorded ladder uses ridge (App. B). Under an affine conditional-mean model, the oracle minimizes expected squared error (Prop. A3); the fitted reference need not. *Surplus* is the median paired log-MSE ratio of the reference to the reader, so positive values favor the reader. Seven frozen checkpoints differ relative to this reference (Table 8): Moirai-2.0-R-small below it even on the linear link (-3.10 $[-3.28, -2.78]$), the Chronos-2 family at the reference with a small surplus on even links, TiRex-2 at the same oscillatory boundary with a large surplus, and the tabular readers above it on every family. Surplus measures relative forecast performance; it does not identify computation outside the affine class.

Degree-1 Hermite share yields median held-out Spearman 0.65 across readers, but weak agreement on TabICL and reversed ordering on TabPFN-TS. Ability slopes span a 237-fold range; one calibration cell fails the registered cross-reader prediction criteria (App. C.2). The held-out-link predictor also fails on TabICL (Fisher-z p=0.009; Fig. 2e).

Limits: lag, noise and changing relations. Chronos-2 reads covariates through lag ℓ=24 and fails by 40 (Fig. 2b). Corrupting the *horizon* covariate segment reduces accuracy, with break-even near one covariate standard deviation (σ^*=1.02–1.22; Fig. 2f), and beyond σ^* harm *grows* to $12.4\times$.

In every tested corruption family, the reader continues to trust a relation verified in context even after that relation breaks at the horizon. A relation *flip* m steps before the horizon is visible in context and motivates truncation; harm is greatest for recent flips (m=16: 2.580; m=256: 0.990; Fig. 4b).

Disjoint seeds reproduce the profile; two real datasets also show covariate use (Table 7).

5 CovLift: Engineering Covariates for Frozen Readers

CovLift compiles covariates for a frozen reader using a fixed basis dictionary, context screen and channel cap. This *input-space reprogramming* (Ye et al., 2025) writes deterministic basis features into covariate channels; Voice2Series learns a transform (Yang et al., 2021). Nonlinear in-context learning depends on learned representations (Sun et al., 2026; Yang et al., 2024; Garg et al., 2022). Our basis exposes poorly read relations for readers near the fitted affine reference (§4).

Inputs and per-window decisions. CovLift-Auto chooses a compiled feed from target and covariate context, a synthetic reader profile, origin, and provider quality available there (Table 9). The offline profile specifies permitted families, lag budget and capacity floor; the context screen selects permitted features. Reliability shrinkage uses the provider's per-site, covariate-only issued-versus-realized error σ_k. Horizon covariates enter forecasting but never transform selection, and horizon targets are never read. *Training-free* excludes gradient training and backbone updates; context regressions still inform decisions.

Each transform addresses a measured deficit (§4). *Basis expansion* appends products $u_i u_j$, squares, absolute deviations and profile-permitted Fourier bases, with u_k standardized from context. *Reliability shrinkage* addresses quality breaks invisible in context; *changepoint truncation* addresses continued trust after a relation reverses and alone shortens the window. *Lag realignment* repairs lags beyond budget, recovering TabICL's tested deficit ($0.996\rightarrow0.106$ at lag 48; Table 15).

Execution. Algorithm 1 specifies CovLift-Auto v5. Its dictionary $\mathcal{D}$ follows the basis order above; $\mathcal{D}_{\mathcal{F}}$ contains its candidates whose families belong to the permitted set $\mathcal{F}$. The total channel cap is $k_{\max}$=12 (2 for the conservative control profile), the support threshold and incremental margin are θ=δ=0.10, and the lag budget b is 24 for Chronos-2 and 4 for TabICL. The noise break-even is $\sigma^\star$; the v5 floor ϕ requires at least one attested natively read nonlinear link. For the screen, y^{det} is the OLS-detrended context target, and R_{nat} centers its argument and removes its ridge fit (10^{-6}) on centered, context-scaled native channels; $f^{\mathrm{res}} = R_{\mathrm{nat}}(f)$. App. C gives the remaining symbol definitions, formulas and guards. We hash the frozen routing decision before forecasting and verify invariance under horizon covariate derangement (Remark A1, App. C). If no transform activates, CovLift returns the byte-identical raw feed.

Algorithm 1 CovLift-Auto v5 on one window. Apply only profile-permitted transforms, freeze the decision, and forecast with unchanged reader weights.

Require: profile $\Pi = (\mathcal{D}, \mathcal{F}, k_{\max}, \theta, \delta, b, \sigma^\star)$, floor ϕ; window $(y^{\mathrm{ctx}}, C^{\mathrm{ctx}}, C^{\mathrm{hor}})$; origin t_0; provider errors σ_k known at t_0

Ensure: forecast $\hat{y}$ of the H steps after t_0; frozen weights, horizon target never read

1: **if** Π invalid, inactive, or its attested nonlinear-link count is missing or below ϕ **then return** $f_{\mathrm{cov}}(y^{\mathrm{ctx}}, C)$ ▷ raw feed

2: $s \leftarrow \arg\max_s \|\hat{\beta}_{[0,s)} - \hat{\beta}_{[s,L)}\|_2$ on a 16-step grid; cut to $[s, L)$ when that statistic exceeds the 0.99 quantile of 32 circular shifts and the statistic is at least 1.5 times that quantile

3: **for all** channels k **do**

4: **if** $b \le 4$, $\hat{\ell}_k > b$, context-fit gain $> 10\%$, and $|\mathrm{ACF}_k(24)| < 0.5 \max_{1\le j\le 8} |\mathrm{ACF}_k(j)|$ **then** shift C_k by $\hat{\ell}_k$

5: **if** $\sigma_k \ge \sigma^\star$ **then** shrink C_k^{hor} toward mean(C_k^{ctx}) by $1/(1+\sigma_k^2)$

6: $r \leftarrow R_{\mathrm{nat}}(y^{\mathrm{det}})$ ▷ context-only screen residual

7: $A \leftarrow$ first $\max(0, k_{\max} - K_0)$ features $f \in \mathcal{D}_{\mathcal{F}}$, dictionary order, with $|f^{\mathrm{res}} r|/(\|f^{\mathrm{res}}\|\,\|y^{\mathrm{det}}\|) \ge \min(1, \max(\theta, \delta))$

8: **if** $A = \emptyset$ and no channel changed **then return** $f_{\mathrm{cov}}(y^{\mathrm{ctx}}, C)$ ▷ abstain, byte-identical

9: **return** $f_{\mathrm{cov}}(y^{\mathrm{ctx}}, [C, A(C)])$

Budget and dictionary. On the certified spring-load grid, Chronos-2's best tested appended budget is K=2 for $L \le 512$ and K=8 at L=1,024. We deploy the four-channel weather-only dictionary at L=1,024; the measured short-context selection-noise effect did not replicate out of sample. Calendar-only squares, absolute deviations and hour-sin/cos products add no covariate information, harm solar (1.08 [1.05, 1.11]), and their removal improves certified load (0.965 [0.938, 0.986]). Every added channel therefore includes a weather covariate.

Versions and scope. Rounds 1–3 use v4 without the capacity gate; Round 4 uses v5. Only expansion activated in confirmed cells. Earlier unconditional realignment failed; App. C.3 records that failure and the later guard variants. *Empirical support for basis expansion.* Across 21 synthetic relations and three basis orders on Chronos-2 (63 points), the gain from unscreened expansion rises with the in-window variance the basis makes linear (Spearman 0.93; post-hoc slope 0.70 [0.59, 0.80] in log-variance units, an association without an identity law; App. D.1). Compiled inputs can still worsen forecasts. We assess harm empirically against the preregistered tolerance (§6).

6 Evaluation

Protocol. We test repairs and transfer across energy, retail and benchmark tasks. Confirmatory geographic transfer requires issued covariate forecasts, contamination immunity from periods provably after the cutoff, verifiable non-use under a locked manifest, and enough series for cluster inference. LTSF/ETT and GIFT-Eval (Aksu et al., 2024) have no future-known channels, while fev-bench-era corpora largely predate the cutoffs (App. C); energy therefore dominates the eligible pool. A development replay falsified unconditional lag realignment (Table 17), after which we re-registered all evaluations.

Table 1: **Preregistered tests on new series.** Compiled/raw geometric MSE ratios (< 1 favors CovLift-Auto), 95% hierarchical-bootstrap CIs, and within-round Holm-adjusted p. n counts series; $H{=}24$ (Round 1) or 96 (Rounds 2–4). Round-2–4 load rows use the legacy day-1 weather feed; timing-corrected analysis has a distinct endpoint (App. D). Classifications retain the original registered analyses: confirmed/replicated meet effect and control criteria; withheld meet the effect criterion with some registered controls unmet. Round 4's shared control criteria withhold confirmation for both readers (§6).

Round	Dataset	Reader (n)	MSE ratio [95% CI]	Holm p	Conclusion
1	Bikeshare	Chronos-2 (2)	$1.046\ [0.978, 1.130]$	1.000	No clear gain
1	Bikeshare	TabICL (2)	$1.036\ [0.979, 1.094]$	1.000	No clear gain
1	Electricity load	TabICL (4)	$0.982\ [0.921, 1.042]$	0.823	No clear gain
2	Electricity load	Chronos-2 (6)	$0.819\ [0.717, 0.915]$	≤ 0.00025	Confirmed
3	Electricity load	Chronos-2 (8)	$0.831\ [0.714, 0.945]$	0.00175	Replicated
4	Electricity load	Chronos-2 (16)	$0.901\ [0.834, 0.969]$	0.00225	Withheld
4	Electricity load	TabICL (16)	$0.937\ [0.905, 0.963]$	≤ 0.00050	Withheld

Table 1 reports primary hypotheses against criteria fixed before execution; a *gain* has an MSE-ratio CI below one. *Confirmation* also requires the preregistered effect test, a 95% CI excluding one on the stated side, and all audits; audit failure withholds a statistically supported effect. *Confirmed harm* has a CI lower bound above 1.05 on the evaluated windows (§5). Real-data results use the context-only pipeline in Table 9, with decision-invariance audits passing throughout (App. C).

Controlled repairs across readers. All four transform families repair targeted deficits on both Chronos-2 checkpoints and TabICL (Table 15); in the four-trap stress test, the v5 gate abstains below its capacity floor after three spurious-covariate threshold exceedances (Table 16).

Preregistered prospective evaluation. For each round, we average paired log-MSE differences over non-overlapping origins within series and then equally across series (a geometric MSE ratio; abstentions count as one). We use hierarchical series-cluster/origin-block bootstrap and within-round Holm correction (Holm, 1979). With 6–16 clusters, we emphasize interval bounds; App. C.3 records deviations and audit disclosures.

Round 1 froze thirteen post-cutoff series at $H{=}24$ with a day-ahead-forecast tolerance pool: *the hypotheses did not confirm, and the pooled tolerance criterion passed* (Table 1).

Generalization to new regions: Rounds 2 and 3. Under the legacy day-1 weather feed, Chronos-2's $H{=}96$ load gain was confirmed on six new grid regions and replicated on the next eight: a preregistered 17–18% geometric-MSE reduction (Table 1). Round 2's pair and horizon followed Round 1; Round 3's regions were fixed ex ante (App. C.3). Shuffling eliminates every gain. SRP was adverse ($1.105\ [1.003, 1.212]$), while the pooled criterion passes with geographic clustering and in the worst leave-one-region-out case (Table 18).

Availability and seasonal transfer. The corrected feed uses forecast references at least six hours before origin; publication timestamps are unobserved. Its spring auto/raw ratio is $0.861\ [0.797, 0.927]$, versus $0.840\ [0.757, 0.916]$ for leaked legacy day-1. The sole endpoint on 210 July–August origins is $0.968\ [0.922, 1.007]$: the spring gain survives this assumed availability rule, while seasonal efficacy remains unestablished (App. D).

Testing transfer across readers: Round 4. On sixteen new regions, Round 4 tested two $H{=}96$ hypotheses (Holm $m{=}2$): a third Chronos-2 and first confirmatory TabICL test. Both met the statistical criterion (Table 1), but one of twelve load cells with CIs below one (WALC$\times$TabICL) fails the shuffle check (0.998 versus the >1.05 requirement); the shared criterion withholds confirmation for both readers (Table 18).

Primary finding: longer context turns harm into benefit. On spring load, Chronos-2's $K{=}8$ expansion against same-length raw has log-ratio contrast $0.388\ [0.290, 0.494]$ (ratio 1.474) from

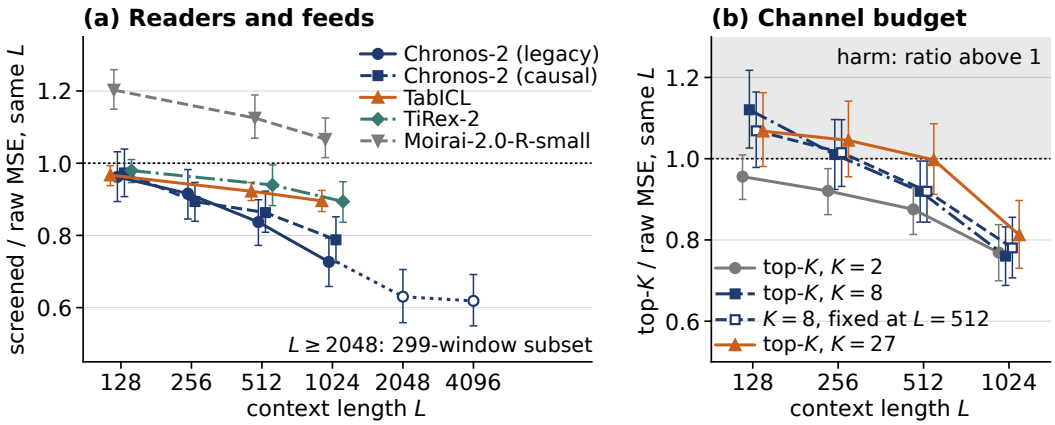

Figure 3: **The capacity regularity on spring load.** (a) Deployed screened-expansion/raw error ratio at common context length L (geometric mean of paired per-window MSE ratios; below 1 favors expansion; 95% moving-block intervals, block 8, 2,000 draws) for four readers on 328 certified windows. Open circles at L=2,048 and 4,096 use the 299-window subset eligible at 4,096; its $L\leq$1,024 curve differs by at most 0.016. (b) Chronos-2 $K\times L$ surface on the legacy feed: top-K appended channels for K=2, 8, 27, plus K=8 identities fixed at L=512 (dashed). The band marks harm.

L=128 to 1,024 (Fig. 3b): 7.5× a stylized analytic least-squares variance reference of 0.052 (5.4× corrected). Excess peaks at 16.5× [11.0, 22.7] for K=2 and falls to 1.7× [1.1, 2.3] for 27 channels (corrected: 11.1×, 1.4×; Table 11).

These preregistered tests use 328 certified spring windows at $L \in \{128, 256, 512, 1{,}024\}$ and the top K of 27 support-ranked products, squares, cubes, quartics, absolute deviations and Fourier terms of three native channels. Every ratio and interval was independently reconstructed from per-window records.

Selection versus context. Fixing K=8 identities at L=512 gives ratios 1.069 versus 1.120 at L=128 and 0.780 versus 0.760 at L=1,024: re-selection hurts at short context and helps at long context on spring load. The registered out-of-sample component test did not confirm (0.038 $[-0.014, 0.090]$ over 799 held-out windows). The best budget moves from K=2 ($L\leq$512) to K=8 (L=1,024) on both feeds; the registered K=27-against-2 contrast did not confirm (0.056 $[-0.041, 0.166]$; the full bank has no selection noise), while the corrected K=8-against-2 interaction confirmed out of sample (0.128 $[0.069, 0.195]$).

Transfer across readers. Readers differ in the preregistered direction (-0.179 $[-0.304, -0.066]$), and their L=512 order matches the linear-reader prediction fixed before TiRex-2 and Moirai-2.0 (App. D.1). The 27-channel bank leaves TabICL unharmed at L=128 (0.937); the deployed screen gives TiRex-2 a smaller Chronos-2-like slope (0.980, 0.940, 0.894) and harms Moirai-2.0 at every length, less with longer context (1.203, 1.125, 1.066).

Longer-context limit. On 299 windows of 13 regions, the Chronos-2 ratio reaches 0.630 and 0.619 at L=2,048 and 4,096 (0.444 $[0.322, 0.593]$), while raw error rises 12% and 20% from L=1,024 (0.18 of the 0.44 contrast). Expansion has its lowest absolute error at 2,048; the gain levels off.

Transfer across populations. On 16 unseen regions, ratios decrease 1.067, 0.898, 0.779 at L=128, 512, 1,024; winter gives 1.074, 0.976, 0.979, and summer 1.013, 0.969, 0.991, with 59% abstention (Table 13, App. D.1). The spring fit failed registered numerical transfer on all three populations. The sign test confirmed on unseen regions (0.350 $[0.238, 0.463]$) and winter (0.177 $[0.036, 0.348]$), but not summer (0.051 $[-0.084, 0.181]$): the regularity is population-dependent, and a transferable scaling law remains unestablished. A known-link synthetic world confirms excess noise-channel cost (3.2× the stylized variance benchmark; App. D.1).

Choosing dictionary and context together. Calendar-derived channels dominate solar selections and increase error (1.083 $[1.055, 1.114]$). The weather-only dictionary is near raw (1.001

$[0.989, 1.014]$) and improves certified load (0.965 $[0.938, 0.986]$); at L=1,024, it beats deployed configuration on both spring feeds (0.845 $[0.797, 0.899]$, 0.901 $[0.841, 0.966]$) and unseen regions (0.841 $[0.794, 0.891]$). Compiled L=1,024/raw L=512 is 0.71 on certified data. Winter and summer point estimates (0.910, 0.941) favor the combined configuration, but registered intervals include one and its advantage is the backbone's own context gain (Table 14, App. D.1).

Generation targets are a boundary. On 1,050 solar and 1,082 wind windows from 12 regions each, with issued cloud-cover and wind-speed covariates, Chronos-2 reads the physical covariate natively (numeric-only/raw error ratio 1.27 on solar, 2.05 on wind) and none of the tested appended channels yields a confirmed gain: the explicit cloud×diurnal interaction (1.011 $[0.984, 1.050]$) and the cube of wind speed (1.008 $[0.988, 1.029]$) fail the registered interaction prediction. Under the deployed expansion both readers reach the same level (numeric-only/expansion error ratio 1.19 for both, on the 265-window TabICL subset); TabICL gains from the explicit interaction (0.952 $[0.922, 0.981]$) while it extracts less from raw (numeric-only/raw error ratio 1.14 against Chronos-2's 1.25 on that subset).

Baselines and harm audits. At zero per-target training cost, compilation is competitive with the best *trained* ChronosX/UniCA-family adapter on one of four valid cells; two intervals span one and the adapter dominates the fourth (App. E). At H=24 under an inference-faithful protocol, TATO (Qiu et al., 2026) has confirmed harm in 11/26 cells; TFMAdapter (Dange & Sarawagi, 2025) and CovLift-Auto have none. TFMAdapter's error ratios exceed one on pools with no measured deficit; screening maintains parity (Table 21).

At H=96 on fourteen Rounds-2/3 regions, TATO increases error (1.325 $[1.049, 1.733]$; CI lower bound below 1.05), TFMAdapter has no confirmed effect (1.013 $[0.990, 1.038]$), and equal-region geometric ratios use L=960. On the same origins at L=512, compilation reduces the geometric MSE ratio by 17–18%, reported by round.

Benchmark breadth. On 23 single-target known-covariate tasks selected from fev-bench (Shchur et al., 2025), CovLift-Auto scores -2.1% $[-4.3, -0.4]$ against raw Chronos-2 (benchmark skill; higher is better). Point estimates worsen on 11 tasks; none meets the per-task confirmed-harm criterion. Keeping compiled inputs only when the reader's own backtest favors them recovers $+0.2\%$ $[-0.1, +0.7]$ (a registered partial; App. D.1). Readers use covariates natively on two electricity-price markets and three ProEnFo load datasets, leaving no measured deficit to repair (App. D.1).

Nonlinearity and seasonal limits. Within load windows where CovLift-Auto acts, improvement increases monotonically with temperature–load nonlinearity; a calendar placebo gives no gradient. The registered nonlinearity-share prediction fails on winter, and ties weaken the registered pooled endpoint (App. D, including all numerical results).

Forecasting versus selection. Compilation beats the full same-input classical panel on all 30 Chronos-2 load cells of Rounds 2–4 (Table 19); no Rounds-1–4 cell meets confirmed harm (CI lower bound >1.05; App. D). A generic control with matched dictionary, screen and cap gives identical reader inputs at all 543 origins, so the comparison cannot identify incremental selection gain; Fig. 3a assesses diagnostic prediction separately.

Retail transfer. A registered retail test transfers a frozen profile-informed selector from H=96 to H=14 (L=512; one chain, 27 stores, eight dates). Its geometric error ratio against numeric-only is 0.813 (18.7% lower), but gains over fixed expansion, rolling validation and generic no-profile selection do not confirm; the all-four criterion fails. It always chooses splines, leaving profiling's incremental value unestablished. Promotion availability is assumed and pretraining exposure unknown (App. E).

7 Conclusion

Across controlled probes, energy, retail and benchmarks, we examine under-generalization in frozen models. CovLift improves unseen-region load forecasts; the retail pipeline improves over numeric-only forecasting. Mixed transfer and unestablished profiling gains bound these results (§6). Diagnosing reader responses and matching features and context can support reliable zero-shot forecasting without retraining. We hope this work contributes to a broader reliability goal, also motivated by hallucination in language models: translating foundation-model capabilities into dependable behavior.

Reproducibility Statement

The artifact contains code, configurations, tests and compact saved results, with offline commands to reconstruct the central context-capacity figure and baseline summaries. It includes the diagnostic interface (`covprobe/`); App. C.3 reports its historical live Chronos-2 validation. Model weights, bulk raw/prepared data and full historical execution caches are omitted. Package hashes certify distributed bytes; retained experimental hashes refer to original executions, with sanitized copies identified as derivatives. Fresh forecast runs require registered inputs, reader-specific environments and new provenance records. Appendix A indexes the evidence. The appendices contain definitions, proofs, controls, capacity results, prospective and falsification summaries, and baseline comparisons.

AI Use Statement

We used large-language-model assistance during research development for code drafting, experiment orchestration, log summarization, figure and table preparation, and prose editing. The research questions, hypotheses, preregistered designs, and interpretations are our own; we inspected and verified all experiments, numerical claims, citations, and final manuscript text, and we take full responsibility for the submission.

Ethics Statement

Forecasting errors can affect energy operations and retail planning. The observed gains depend on the reader, context and population; our empirical harm criteria provide no deployment guarantee. Applications require validation under local data availability and distribution shifts, and compliance with dataset, code and model licenses. We report adverse and inconclusive results, protocol deviations and forecast-availability limitations to support responsible interpretation.

References

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

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations (ICLR)*, 2023.

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 (TMLR)*, 2024.

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.

Gaspard Berthelier, Mariia Baranova, Andrei-Tiberiu Pantea, Etienne Le Naour, Adrien Petralia, Tahar Nabil, and Themis Palpanas. Investigating simple target-covariate relationships for Chronos-2 and TabPFN-TS. *arXiv preprint arXiv:2605.12200*, 2026.

Satwik Bhattamishra, Arkil Patel, Phil Blunsom, and Varun Kanade. Understanding in-context learning in transformers and LLMs by learning to learn discrete functions. In *International Conference on Learning Representations (ICLR)*, 2024.

Blake Bordelon, Mary Letey, and Cengiz Pehlevan. Theory of scaling laws for in-context regression: Depth, width, context and time. In *International Conference on Learning Representations (ICLR)*, 2026.

Andrea Carriero, Davide Pettenuzzo, and Shubhranshu Shekhar. MacroCast: A vintage-consistent time series foundation model for real-time macroeconomic forecasting. *arXiv preprint arXiv:2606.28670*, 2026.

Wei Chen and Yuxuan Liang. Learning with calibration: Exploring test-time computing of spatio-temporal forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Hanyin Cheng, Xingjian Wu, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. CoRA: Boosting time series foundation models for multivariate forecasting through correlation-aware adapter. In *International Conference on Learning Representations (ICLR)*, 2026.

Afrin Dange and Sunita Sarawagi. TFMAdapter: Lightweight instance-level adaptation of foundation models for forecasting with covariates. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 10148–10167, 2024.

Sofiane Ennadir, Tianze Wang, Oleg Smirnov, Sahar Asadi, and Lele Cao. Be wary of your time series preprocessing. *arXiv preprint arXiv:2602.17568*, 2026.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-ts: A general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7522–7529, 2023.

Everette S. Gardner and Ed McKenzie. Forecasting trends in time series. *Management Science*, 31(10): 1237–1246, 1985. doi: 10.1287/mnsc.31.10.1237.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Lu Han, Han-Jia Ye, and De-Chuan Zhan. SIN: Selective and interpretable normalization for long-term time series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 17437–17453, 2024.

Lu Han, Yu Liu, Lan Li, Qiwen Deng, Jian Jiang, Yinbo Sun, Zhe Yu, Binfeng Wang, Xingyu Lu, Lintao Ma, Han-Jia Ye, and De-Chuan Zhan. UniCA: Unified covariate adaptation for time series foundation model. In *International Conference on Learning Representations (ICLR)*, 2026.

Noah Hollmann, Samuel Müller, and Frank Hutter. Large language models for automated data science: Introducing CAAFE for context-aware automated feature engineering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2): 65–70, 1979.

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 (TMLR)*, 2026. ISSN 2835-8856.

Ao Hu, Liangjian Wen, Jiang Duan, Yong Dai, He Yan, Dongkai Wang, Jun Wang, Yukun Zhang, Ruoxi Jiang, and Zenglin Xu. PMDformer: Patch-mean decoupling information transformer for long-term forecasting. In *International Conference on Learning Representations (ICLR)*, 2026a.

Hao Hu and Xue-shan Ai. Exogenous dropout: A simple, strong baseline for corruption-robust time series forecasting with covariates. *arXiv preprint arXiv:2607.05452*, 2026.

Yifan Hu, Jie Yang, Tian Zhou, Peiyuan Liu, Yujin Tang, Rong Jin, and Liang Sun. Bridging past and future: Distribution-aware alignment for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2026b.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 3rd edition, 2021.

Jeonghwan Im and Hyuk-Yoon Kwon. COSA: Context-aware output-space adapter for test-time adaptation in time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2026.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025.

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

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989. doi: 10.1214/aos/1176347265.

Daojun Liang, Qi Li, Yinglong Wang, Jing Chen, Hu Zhang, Xiaoxiao Cui, Qizheng Wang, and Shuo Li. The forecast after the forecast: A post-processing shift in time series. In *International Conference on Learning Representations (ICLR)*, 2026.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 32369–32399, 2024.

Zhiding Liu, Mingyue Cheng, Zhi Li, Zhenya Huang, Qi Liu, Yanhu Xie, and Enhong Chen. Adaptive normalization for non-stationary time series forecasting: A temporal slice perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14273–14292, 2023.

Jiecheng Lu, Xu Han, and Shihao Yang. ARM: Refining multivariate forecasting with adaptive temporal-contextual learning. In *International Conference on Learning Representations (ICLR)*, 2024.

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, 2025. v3 title (Feb 2026).

Marcel Meyer, Sascha Kaltenpoth, Henrik Albers, Kevin Zalipski, and Oliver Müller. Ts-arena – a live forecast pre-registration platform. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26), Datasets and Benchmarks Track*, pp. 9558–9568, 2026. doi: 10.1145/3770855.3817515.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. TS-RAG: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

OpenSTEF. Dataset card for Liander 2024 short term energy forecasting benchmark. Hugging Face dataset card, 2025.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2020.

Manya Pandey, Dhruv Kumar, Murari Mandal, and Saurabh Deshpande. GITCO: Gated inference-time context optimization in TSFMs. *arXiv preprint arXiv:2606.05332*, 2026.

Sebastian Pineda Arango, Pedro Mercado, Shubham Kapoor, Abdul Fatir Ansari, Lorenzo Stella, Huibin Shen, Hugo Henri Joseph Senetaire, Ali Caner Turkmen, Oleksandr Shchur, Danielle C. Maddix, Michael Bohlke-Schneider, Bernie Wang, and Syama Sundar Rangapuram. ChronosX: Adapting pretrained time series models with exogenous variables. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *PMLR*, 2025.

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

Zhen Qin, Jiachen Jiang, and Zhihui Zhu. Learning to adapt: In-context learning beyond stationarity. In *International Conference on Learning Representations (ICLR)*, 2026.

Yunzhong Qiu, Zhiyao Cen, Zhongyi Pei, Chen Wang, and Jianmin Wang. Adapt data to model: Adaptive transformation optimization for domain-shared time series foundation models. In *International Conference on Learning Representations (ICLR)*, 2026.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 50817–50847, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. To the cutoff… and beyond? a longitudinal perspective on LLM data contamination. In *International Conference on Learning Representations (ICLR)*, 2024.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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*, 2025.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Haoyuan Sun, Ali Jadbabaie, and Navid Azizan. On the role of transformer feed-forward layers in nonlinear in-context learning. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory (ALT)*, volume 313 of *PMLR*, 2026.

William Toner and Luke Darlow. An analysis of linear time series forecasting models. In *International Conference on Machine Learning (ICML)*, 2024.

Fangxin Wang, Ziyi Zhang, Diyi Zhuang, Langzhou He, Shiyu Wang, Baichuan Mo, and Philip S. Yu. When do corrective features help? an agent for corrective feature discovery on black-box forecasters. *arXiv preprint arXiv:2608.05207*, 2026a.

Hao Wang, Licheng Pan, Yuan Lu, Zhixuan Chu, Xiaoxi Li, Shuting He, Zhichao Chen, Qingsong Wen, Haoxuan Li, and Zhouchen Lin. DistDF: Time-series forecasting needs joint-distribution Wasserstein alignment. In *International Conference on Learning Representations (ICLR)*, 2026b.

Qixun Wang, Yifei Wang, Xianghua Ying, and Yisen Wang. Can in-context learning really generalize to out-of-distribution tasks? In *International Conference on Learning Representations (ICLR)*, 2025.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Zhijie Wang, Bo Jiang, and Shuai Li. In-context learning on function classes unveiled for transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 50726–50745, 2024b.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh Agrawal, Sandeep Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations (ICLR)*, 2025.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 53140–53164, 2024.

Chao-Han Huck Yang, Yun-Yun Tsai, and Pin-Yu Chen. Voice2Series: Reprogramming acoustic models for time series classification. In *International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 11808–11819, 2021.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: Contextual generalization of trained transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yanfeng Yang, Siwei Chen, Pingping Hu, Zhaotong Shen, Yingjie Zhang, Zhuoran Sun, Shuai Li, Ziqi Chen, and Kenji Fukumizu. Conditionally whitened generative models for probabilistic time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2026.

Zesheng Ye, Chengyi Cai, Ruijiang Dong, Jianzhong Qi, Lei Feng, Pin-Yu Chen, and Feng Liu. Neural network reprogrammability: A unified theme on model reprogramming, prompt tuning, and prompt instruction. *arXiv preprint arXiv:2506.04650*, 2025.

Qiuhao Zeng, Changjian Shui, Long-Kai Huang, Peng Liu, Xi Chen, Charles Ling, and Boyu Wang. Latent trajectory learning for limited timestamps under distribution shift over time. In *International Conference on Learning Representations (ICLR)*, 2024.

Tianping Zhang, Zheyu Zhang, Zhiyuan Fan, Haoyan Luo, Fengyuan Liu, Qian Liu, Wei Cao, and Jian Li. OpenFE: Automated feature generation with expert-level performance. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023.

Xu Zhang, Peng Wang, and Wei Wang. Lost in the non-convex loss landscape: How to fine-tune the large time series model? In *International Conference on Learning Representations (ICLR)*, 2026.

Zhiyuan Zhao, Haoxin Liu, and B. Aditya Prakash. Tackling time-series forecasting generalization via mitigating concept drift. In *International Conference on Learning Representations (ICLR)*, 2026.

Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, et al. General scales unlock AI evaluation with explanatory and predictive power. *Nature*, 2026.

A Evidence guide

The appendices supply the definitions and evidence needed to assess the paper's claims. Appendix B gives the mathematical references and numeric controls; App. C gives the diagnostic measurements, implementation contract and protocol disclosures; App. D reports the context and transfer tests; App. E reports component repairs, prospective audit boundaries, comparators and the retail selection test. Primary prospective decisions appear in Table 1. We consolidate secondary per-cell records into summaries with their original estimators and retain failed hypotheses and consequential limitations. No separate document is required to follow these references.

B Mathematical references and numeric controls

B.1 Normalization and affine-equivariant forecasts

Proposition A1 (Exact normalization)**.** *For a nonconstant context, let μ and $\sigma > 0$ be its exact mean and standard deviation, and $N(\mathbf{x}) = (\mathbf{x} - \mu)/\sigma$, with population standard deviation. Context indices are $t = 1, \dots, L$; bars denote context averages, a is level, b slope, and η_t the residual. (a) $N(c\mathbf{x} + d) = N(\mathbf{x})$ for $c > 0$. (b) If $x_t = a + bt + \eta_t$, normalization retains $[b(t - \bar{t}) + \eta_t - \bar{\eta}]/\sigma$. For noiseless $b \neq 0$, this equals $\operatorname{sgn}(b)(t - \bar{t})/s_L$, where $s_L^2 = (L^2 - 1)/12$.*

Proof. Under $\mathbf{x} \mapsto c\mathbf{x} + d$, the mean becomes $c\mu + d$ and the standard deviation becomes $c\sigma$, proving (a). Subtracting the mean in (b) removes $a + b\bar{t}$; without noise, $\sigma = |b| s_L$. □

These are properties of exact RevIN (Kim et al., 2022); implementation details can alter them. The translation invariance and inverse-scale dependence also appear in the expressivity analysis of Ennadir et al. (2026). Trend shape survives: a linear predictor can extrapolate the normalized noiseless ramp exactly.

The read-only-through-the-normalizer assumption (A1) permits a frozen backbone ϕ and any measurable correction Δ depending only on its normalized input and activations:

$$g(\mathbf{x}) = \sigma(\mathbf{x}) \left[\phi(N(\mathbf{x})) + \Delta\big(N(\mathbf{x}), \{\phi_\ell(N(\mathbf{x}))\}_\ell\big)\right] + \mu(\mathbf{x}). \tag{1}$$

Here ϕ_ℓ is the activation at layer ℓ. The class $\mathcal{I}$ consists of these normalized-context readouts followed by inverse normalization; each is affine-equivariant. The backbone symbol ϕ is local to this construction; Algorithm 1 uses ϕ for its capacity floor.

Proposition A2 (Affine-orbit factorization)**.** *Let $F : \{\mathbf{x} \in \mathbb{R}^L : \sigma(\mathbf{x}) > 0\} \to \mathbb{R}^H$ satisfy $F(c\mathbf{x} + d) = cF(\mathbf{x}) + d$ for all $c > 0, d \in \mathbb{R}$. (a) There exists G such that*

$$F(\mathbf{x}) = \sigma(\mathbf{x})G(N(\mathbf{x})) + \mu(\mathbf{x}). \tag{2}$$

(b) With covariates $\mathbf{C}$ held fixed under target transformations, $F(\mathbf{x}, \mathbf{C}) = \sigma(\mathbf{x})G(N(\mathbf{x}), \mathbf{C}) + \mu(\mathbf{x})$ is target-affine-equivariant for any G.

Proof. (a) Apply equivariance to $\mathbf{x} = \sigma(\mathbf{x})N(\mathbf{x}) + \mu(\mathbf{x})$ and restrict F to normalized contexts to define G. (b) Substitute the transformed mean and standard deviation and use Proposition A1(a). □

The full normalized window, including residual shape, enters G; the trend coefficient $\tau = bs_L/\sigma$ need not be sufficient. Neither proposition orders forecast errors. For the narrower model $y_h = a_0 + b_{\text{fut}}h + \epsilon_h$, with context-determined anchor a_0 and estimated slope $\hat{b}$, future slope b_{fut}, finite second moments and $\mathbb{E}[\epsilon_h \mid \mathbf{x}, b_{\text{fut}}] = 0$, the same-anchor forecasts $\hat{y}_h = a_0 + \hat{b}h$ and $y_h^{\text{oracle}} = a_0 + b_{\text{fut}}h$ satisfy, for $h = 1, \ldots, H$ and $\text{MSE}(v) = H^{-1}\|\mathbf{y} - v\|_2^2$,

$$\mathbb{E}[\text{MSE}(\hat{\mathbf{y}}) - \text{MSE}(\mathbf{y}^{\text{oracle}}) \mid \mathbf{x}, b_{\text{fut}}] = (b_{\text{fut}} - \hat{b})^2 H^{-1} \sum_{h=1}^{H} h^2. \tag{3}$$

Expansion cancels the conditional cross-term. This identity does not require an equivariant reader to make a slope error or make drift optimal within $\mathcal{I}$.

B.2 Fitted linear reference and scoring definitions

For each window, center and scale time on the context and form $X_c = [\mathbf{1}, t_{1:L}, C_{1:L}]$, $X_h = [\mathbf{1}, t_{L+1:L+H}, C_{L+1:L+H}]$. Here t is standardized time, C the raw supplied covariate matrix, and $\mathbf{1}$ the intercept column: $X_c \in \mathbb{R}^{L\times(K_0+2)}$, $X_h \in \mathbb{R}^{H\times(K_0+2)}$ and $\hat{\beta} \in \mathbb{R}^{K_0+2}$. The fitted reference predicts $X_h\hat{\beta}$ from context regression.

Proposition A3 (Oracle risk under an affine conditional mean)**.** *Let $\mathcal{J}$ be the sigma-field generated by target/covariate context and horizon covariates, so X_h is $\mathcal{J}$-measurable. If the target vector $\mathbf{y} = X_h\beta + \varepsilon \in \mathbb{R}^H$ for fixed β, with $\mathbb{E}[\varepsilon \mid \mathcal{J}] = 0$ and finite second moments, $m = X_h\beta$ minimizes expected squared error among $\mathcal{J}$-measurable forecasts with finite risk: $\mathbb{E}\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \mathbb{E}\|\mathbf{y} - m\|_2^2 + \mathbb{E}\|\hat{\mathbf{y}} - m\|_2^2$.*

Proof. Expand $\mathbf{y} - \hat{\mathbf{y}} = \varepsilon + (m - \hat{\mathbf{y}})$; conditional centering makes the cross-term zero. □

Finite-context regression has no oracle guarantee; asymptotic equality requires $\mathbb{E}\|X_h(\hat{\beta} - \beta)\|^2 \to 0$, an additional prediction-consistency assumption. The recorded reference uses ridge $\lambda = 1$ with a free intercept. In 12 audited Chronos-2 cells its absolute cell-median log-MSE contrast with OLS is below 6×10^{-4}; this aggregate agreement does not bound individual windows. On nonlinear links the affine conditional-mean assumption fails. The ladder's surplus $u_R = \text{median}_w \log[\text{MSE}(\text{lin}, w)/\text{MSE}(R, w)]$ measures a paired empirical comparison, not expected risk or nonlinear computation alone: estimation, regularization and temporal information can affect it. The separate $\log(\text{tct}_{\text{lin}}/\text{tct}_R)$ compares marginal medians and need not equal u_R.

Battery-specific denominators. Numeric-shift skill pools MSE within dataset–seed groups, selects the better of last-value and drift, then averages group ratios. Link/lag/degree, horizon, noise-drift and channel-count profiles instead use a per-window minimum over drift, fitted-damping drift, fixed damping $0.9/0.95/0.99$, and seasonal-naive (last-value for period one); the earlier dose battery uses only drift and fitted-damping drift. These skills add 10^{-9} to baseline MSE. This varying denominator generally does not cancel in FM-own ratios of marginal median skills; it cancels in paired within-window log-skill ratios on positive finite pairs. The prospective scorer averages $\log(\text{skill}_{\text{compiled}} + 10^{-12}) - \log(\text{skill}_{\text{raw}} + 10^{-12})$ within series, then equally across series. The

Table 2: Numeric-shift evidence: eight datasets, five seeds, largest magnitude. Skill averages 40 dataset–seed pooled-MSE ratios against the better of last-value and drift. T/RL compares TSFM MSE with RevIN+linear. Lower is better.

Backbone	Level skill	Scale skill	Trend skill	T/RL
Chronos-Bolt-small	0.43	0.43	1.06	1.24
Chronos-Bolt-base	0.45	0.44	1.03	1.24
Chronos-T5-small	0.54	0.58	2.59	3.09
Chronos-T5-base	0.50	0.53	1.99	2.30
Moirai-1.1-R-small	0.69	0.68	2.25	3.04
Moirai-1.1-R-base	0.65	0.66	2.12	2.84
Moirai-1.1-R-large	0.66	0.66	2.31	2.95
Moirai-MoE-small	0.67	0.68	3.05	3.92
TimesFM-2.0-500m	0.37	0.37	1.26	1.43

context sweep averages paired $\log(\mathrm{MSE} + \epsilon_w)$ differences using saved window guards. Both exponentiate their mean; neither is an arithmetic-mean MSE ratio, and the prospective additive guard prevents exact cancellation to an unregularized MSE ratio.

B.3 Numeric controls and limits of the trend evidence

We use Exchange, ETTh1, ETTh2, ETTm1, Weather, Electricity, Traffic and Solar. Each channel is standardized using training-split statistics. ETT uses its fixed chronological splits (hourly: $8640/2880/2880$ train/validation/test observations; minute: $34560/11520/11520$); the other series use $60/20/20\%$ chronological splits. The shift runner samples channel–origin pairs without replacement within each evaluation seed from eligible test origins, requiring complete context and horizon; paired clean and shifted arms use the same windows. The breadth launch uses seeds 42–46 and up to 48 windows per dataset–seed group. Sampling backbones use eight forecast draws, reduced to their median; the deterministic Bolt models do not use this sampling parameter.

On standardized windows ($L = 512, H = 96$), level adds m to context and target ($m = 1, 2, 4, 8$); scale multiplies both $(2, 4, 8)$; trend adds mi and $m(L + j)$ $(0.002, 0.005, 0.01, 0.02)$. Outliers add random signed spikes to 5% of context $(3, 5)$. Variance rescales the second context half and target about the context mean $(2, 4)$. Regime shifts the first half's mean by m and shrinks its spread to 0.3 $(2, 4)$. Outlier and regime horizons remain clean.

Degradation is shifted/clean skill; recovery is fixed/shifted skill. Level/scale degradation is approximately one; Chronos-T5 level degradation reaches 0.90, so the same-normalization interpretation of T/RL applies to the other seven models. With frozen Timer weights and data, disabling RevIN raises level degradation at magnitude eight from 1.00 to 83.89, isolating normalization's absorption effect. RevIN mitigates trend degradation (12.64 on versus 181.90 off at 0.02), while a stronger linear predictor uses the same normalization (Table 2). Thus trend error is not established as a consequence of normalization. The outlier toggle is architecture-dependent: RevIN helps Timer (1.20 versus 1.30) but hurts the from-scratch control (2.02 versus 1.44). Both controls use the eight datasets and $H = 96$, but Timer uses $L = 480$ and evaluation seeds 42–46, whereas the from-scratch control uses $L = 512$ and evaluation seeds 42–43. The outlier doses also differ: Timer averages magnitudes 3 and 5, while the from-scratch value uses magnitude 4. These are within-control on/off contrasts, not a matched-dose comparison between architectures. The from-scratch control uses one training realization per normalizer and dataset under a common training recipe; its evaluation seeds are window samples, not independent training replicates. Only Timer toggles normalization while holding the pretrained weights fixed.

Natural trend and finite comparisons. Across eight datasets and three seeds (7200 windows/backbone), the nine models have median skill 0.24–0.78 at $R^2 < 0.3$, but 1.04–2.26 in reported bins spanning $0.6 \le R^2 < 0.95$, against the per-window last/drift floor. Near-exact-baseline windows with $R^2 \ge 0.95$ are excluded from that summary. The separate single-seed inference cache has only 157 trend windows, concentrated in Exchange (106) and Weather (37): the directional sign holds in 17/18 backbone–dataset series, but dependence-correct tests are significant in only 1/6

adequately supported series; the two-cluster bootstrap gives $p \approx 0.49$. Natural trend is descriptive corroboration; controlled shifts support the measured claim. Detrending can harm natural trend (TimesFM $1.14 \to 2.22$, Moirai-large $1.76 \to 2.72$). Matched injected-shift fixes also fail under wrong mechanisms: detrending a regime splice causes $8.5\times$ degradation, and local Hampel filtering hurts contiguous bursts (median recovery 1.02, dataset-block 95% CI $[1.01, 1.05]$).

A tabular TabICL arm (3000 windows, five datasets, one seed) likewise wins when flat (0.22) and loses on trend (1.1–1.7), corroborating the deficit in a separate reader arm. These finite model and adaptation comparisons establish no information bound or universal accuracy ceiling.

C Profile measurements and implementation

C.1 Calibration, extraction controls and limits

SNR is the driver-to-background-plus-noise variance ratio over the full context and horizon span. The profile measures the FM-own error ratio $t_{\text{cov}}/t_{\text{num}}$; it does not predict absolute MSE. Table 3 gives the calibration values. The seven-link calibration transfers to eight preregistered held-out links on Chronos-2 (sixteen reader–link predictions committed before either suite): Spearman 0.952, exact permutation $p = 5.7 \times 10^{-4}$, MAE 0.046 versus 0.037 in sample. TabICL is a null ($\rho = 0.19$, $p = 0.33$). The quartic is a visible miss for Chronos-2 (predicted 0.78, measured 0.94) and harms TabICL (2.368 $[1.91, 2.83]$). The sawtooth that is useless to Chronos-2 (1.016) is readable by TabICL (0.188). Thus profile prediction is reader-specific; it does not establish better transform selection.

Table 3: **The capacity profile: exact values for Fig. 2** ($t_{\text{cov}}/t_{\text{num}}$; lower is better). Link calibration: n=400, SNR 2.0; lag calibration: linear link, AR(0.9) covariate. **Bold** × = no extraction; "—" = not evaluated; "none" = no extraction (0.985–0.998); brackets: 95% moving-block CIs.

link g (ℓ=0)	Chronos-2	C2-synth (decont.)	TabICL
linear	0.068 [.060, .081]	0.073 [.061, .086]	0.106 [.094, .119]
tanh	0.176	0.129	0.240
step	0.311	0.295	0.161
cubic	0.440	0.342	0.239
quad	0.722	0.612	0.190
abs	0.815	0.717	0.127
sin	**1.017** ×	**0.997** ×	0.142
interact. e_1e_2	**0.997** ×	—	0.470

lag ℓ (linear g)	Chronos-2	TabICL
4	0.100	none
12	0.131	none
24	0.341	none
32	0.566 [.53, .62]	—
40	**1.027** ×	—
48	**1.096** ×	none

Lag interpretation. The ℓ=4 cell at 0.906 [.888, .947] passes the automated gate (derangement raises it to 0.969, more than the 0.05 margin and still below one) and is also reported *none*, since the AR(0.9) covariate retains ρ_{eff}=0.656 of the lagged value on its own against 0.28/0.08/0.006 at ℓ=12/24/48, so the gain is AR carry-over.
Derangement controls. Every extracted cell crosses one under derangement, with a margin that varies (Chronos-2 1.04–1.46; C2-synth 1.07–1.46; TabICL 1.01–1.46, the low end being its interaction cell). The no-extraction cells run 0.97–1.17, so several sit visibly above one (Chronos-2 interaction 1.17, ℓ=48 1.13).

Lower signal-to-noise check. The seven-link ordering is unchanged at SNR 0.5 on Chronos-2 ($L = 512$, $H = 96$, AR coefficient 0.9, 200 windows for each of seeds 42 and 43). In the order linear,

tanh, step, cubic, quad, abs and sin, the FM-own ratios are 0.233, 0.367, 0.520, 0.612, 0.872, 0.999, 1.058. This is an ordering result: abs no longer has a CI-separated extraction gain, and the interaction was not part of this low-SNR sweep. Table 3 separately reports the synthetic-only checkpoint at SNR 2.0; decontamination does not imply invariance to noise.

Held-out prediction design. Each held-out link uses $L = 512$, $H = 96$, SNR 2.0, an AR(0.9) covariate and 200 windows per seed (42 and 43; 400 total per link and reader). For a link function g, probabilist–Hermite polynomials He_d of degree d and $Z \sim N(0,1)$, define $v_d(g) = \mathbb{E}[g(Z)\mathrm{He}_d(Z)]^2/d!$. For reader R, let $q_R(g)$ be its FM-own ratio on link g and $q_{R,d}$ its measured ratio on He_d. With normalized spectral weights $w_d = v_d / \sum_{j=1}^{6} v_j$, the predictor is:

$$\widehat{q}_R(g) = \frac{\sum_{d=1}^{6} v_d(g) q_{R,d}}{\sum_{d=1}^{6} v_d(g) + 10^{-12}}.$$

The implementation estimates the spectrum with 2,000,000 standard-normal draws, seed zero, after centering the link; no held-out response enters the prediction. Table 4 separates the zero-parameter curve from comparators fitted to the held-out responses and from the raw-correlation comparator fitted only on calibration cells.

Table 4: **Out-of-sample link readability: all eight held-out links.** Reader-specific predictions combine each link's Hermite spectrum with its reader's separately measured He_d curve. They were committed before either suite ran, with no fitting to held-out responses. Rows follow the Chronos-2 prediction.

	Chronos-2		TabICL	
link	predicted	measured [CI]	predicted	measured [CI]
softsign	0.118	0.132 [0.12, 0.15]	0.104	0.180 [0.16, 0.20]
logistic	0.122	0.132 [0.12, 0.15]	0.105	0.196 [0.18, 0.22]
hard_clip	0.143	0.155 [0.13, 0.17]	0.109	0.192 [0.18, 0.22]
cube_root	0.144	0.182 [0.17, 0.20]	0.108	0.183 [0.17, 0.20]
gauss_bump	0.771	0.697 [0.65, 0.75]	0.254	0.127 [0.11, 0.16]
quartic	0.777	0.940 [0.91, 0.96]	0.266	2.368 [1.91, 2.83]
sawtooth	0.979	1.016 [0.98, 1.06]	0.184	0.188 [0.17, 0.21]
he_mix	1.016	0.996 [0.94, 1.03]	0.170	0.185 [0.17, 0.20]

Prediction accuracy. Chronos-2: Spearman 0.952 (exact one-sided permutation $p = 0.0006$, $n = 8$), MAE 0.046. TabICL: 0.190 ($p = 0.33$), MAE 0.322; the ranking transfers on Chronos-2 but is not established for TabICL. Separately, the measured central spread differs: IQR 0.824 (Chronos-2) vs 0.012 (TabICL). TabICL's responses cluster together except for the quartic outlier.
Effective degree. A reader-free effective-degree statistic $\sum_d d\, w_d$ matches the Chronos-2 ranking (ρ=0.952, p=0.0006). The measured curve improves numerical prediction: MAE 0.046, versus 0.136 for the statistic after a best-fit affine rescaling to the same held-out responses (two fitted parameters versus zero).
Context screens. Screen statistics scored the same way: the raw covariate's detrended linear correlation with the target context reaches MAE 0.048 fitted on these eight responses, 0.050 with its two parameters fitted only on the 13 calibration cells, and Spearman -0.976; screens over the shipped dictionary (0.298) and the extended dictionary (0.155) do not predict the level, and their rank agreement carries the opposite sign (0.667, 0.738).

Smoothness is part of the regime. Holding the marginal, SNR and L/H fixed, the pre-stated smoothness test changes the apparent degree cutoff and scrambles link order (Table 5). This withdraws the polynomial-kernel mechanism interpretation: the curve describes the reader–window system. The He_2–He_4 CIs at $\varphi \le 0.6$ are disjoint from the frozen curve. At $\varphi = 0.99$ the ladder is non-monotone; link-order Spearman is 0.75, versus 1.0 at $\varphi \in \{0.3, 0.6\}$.

Testbeds and controls. The synthetic event process controls salience and SNR. EPF (2011–2018) lies in Chronos-2's pretraining era, so we compare public and synthetic-only weights (Table 6).

Table 5: **Smoothness test.** Chronos-2 FM-own ratios (lower is better). Extraction requires CI upper < 1 and derangement collapse. Link rows use 400 windows, seeds 42–43; the frozen $\varphi = 0.9$ degree ladder uses 900, seeds 0–2, and swept ladders use 400, seeds 42–43. The cutoff shift also separates matched-size swept columns.

φ	0	0.3	0.6	0.9	0.99
CI-crossing cutoff d^*	2	2	2	4	6
He_2	0.948	0.932	0.905	0.737	0.637
He_3	1.022	1.027	1.021	0.836	0.605
He_4	1.029	1.032	1.015	0.893	0.700
step	0.439	0.373	0.333	0.311	0.787
abs	0.931	0.971	0.933	0.815	0.598
quadratic	0.947	0.936	0.910	0.722	0.675

Table 6: **Decontamination: covariate extraction replicates on a synthetic-only checkpoint.** The identical EPF protocol on the public `amazon/chronos-2` weights vs. `chronos-2-synth`, trained on *synthetic data only*. The covariate skill t_{cov} agrees market-for-market to within 8%; the numeric base t_{num} is only 19–38% stronger on the public weights (PJM 19%, DE 20%, NP 22%, BE 34%, FR 38%), and *both* sit well under the trivial baseline ($\ll 1$). Lower is better.

	t_{num} (numeric base)		t_{cov} (covariate)	
market	public	synth-only	public	synth-only
NP	0.290	0.373	0.202	0.207
DE	0.305	0.380	0.166	0.158
BE	0.267	0.405	0.190	0.204
FR	0.223	0.360	0.137	0.128
PJM	0.166	0.205	0.114	0.115

Freshcov uses five cities' 2025–2026 Open-Meteo forecasts for ERA5 temperature, after the training cutoff. Its extraction is therefore post-cutoff; this does not remove the separate forecast-vintage qualification for the prospective load studies. Table 7 reports the FM-own ratio of median baseline-normalized skills, with moving-block 95% CIs. Every real-covariate CI is below one; calendar placebos are near one, and cross-window horizon derangement destroys the gain while leaving the numeric path untouched. These controls identify use of the specific covariate.

C.2 Reader ladder and failed cross-reader prediction

Demand and ability. In each leave-one-reader-out fold, a demand candidate is chosen on the six training readers before held-out scoring. The degree-1 Hermite share w_1 wins every fold (fold MAE 0.090–0.115 in FM-own ratio units versus 0.13–0.19 for odd/even shares). A held-out reader receives the mean training slope and one linear-link anchor. None meets both registered bars, Spearman ≥ 0.8 and MAE ≤ 0.10: Chronos-2 $(0.75, 0.50)$, TiRex-2 $(0.74, 0.27)$, TabPFN $(0.71, 0.20)$, synthetic-only Chronos-2 $(0.65, 0.48)$, Moirai-2 $(0.32, 1.52)$, TabICL $(0.25, 0.19)$ and TabPFN-TS $(-0.26, 0.08)$. Coverage differs by reader. Median held-out rank correlation is 0.65, but own slopes span -2.08 to $+0.01$ (a 237-fold range). Ordering does not yield calibrated cross-reader prediction.

Dynamics and channel use. Horizon-only noise breaks even at $\sigma^* = 1.02$–1.22 covariate standard deviations, then harms up to $12.4\times$. For a relation flip m steps before the horizon, Chronos-2's FM-own ratio is 2.580 at $m = 16$ and 0.990 at $m = 256$; its best tested ridge reference uses full context, while TabICL's flip reference has a roughly 64-step half-life. A relation ending within context favors finite forgetting (half-life 128 on the two Chronos-2 checkpoints, 16 on TabICL). Horizon-only relation breaks evade the tested context detectors: backtest correlations with horizon skill are Pearson 0.003 / Spearman 0.097 on 1,500 windows. Their mean harm is $15.4\times$, versus $1.7\times$ when the covariate is wrong throughout.

Table 7: A frozen Chronos-2 extracts a future-known exogenous covariate (exact values and 95% CIs). FM-own ratio $t_{\mathrm{cov}}/t_{\mathrm{num}}$ of median baseline-normalized skills; <1 means the covariate helps *beyond* the model's own numeric forecast, and lower is better. Every real-covariate 95% CI (moving-block bootstrap over windows) lies below one, and a zero-information *calendar* covariate is inert (≈ 1). *Shuffling* the covariate across windows (leak-negative) collapses the gain, and on real data inverts it.

Testbed	unit	$t_{\mathrm{cov}}/t_{\mathrm{num}}$ [95% CI] (real covariate)	calendar placebo	shuffled (leak)
freshcov post-cutoff, immune	Sydney	0.029 [0.021, 0.035]	1.07	12.3
	Chicago	0.101 [0.080, 0.124]	1.15	3.5
	London	0.127 [0.103, 0.170]	1.04	7.1
	Berlin	0.177 [0.147, 0.215]	1.04	8.4
	Madrid	0.186 [0.142, 0.243]	0.99	11.3
EPF real, pretrain-era	DE	0.545 [0.478, 0.653]	1.14	3.2
	FR	0.614 [0.508, 0.718]	0.97	5.7
	PJM	0.687 [0.587, 0.815]	1.09	3.5
	NP	0.697 [0.593, 0.822]	1.11	4.8
	BE	0.711 [0.606, 0.812]	0.93	4.8
synthetic	event DGP	0.561 [0.514, 0.601]	n/a	1.03

Table 8: **The linear-reader ladder.** Surplus $u = \mathrm{median}_w \log[\mathrm{MSE}(\mathrm{lin}, w)/\mathrm{MSE}(R, w)]$ compares each frozen reader with the recorded per-window ridge reference; $u > 0$ favors the reader and $u < 0$ favors the reference. Columns give the magnitude at the reference on monotone links, the even-link surplus, the odd-cubic surplus, the sin surplus, and the ladder score over the four shared cells.

reader	cells	at ref. $\lvert u\rvert$	even	odd-cubic	sin	ladder
TabPFN-TS	7	0.884	2.415	1.235	2.262	2.082
TabPFN	13	0.453	1.206	0.818	1.332	1.137
TabICL	18	0.534	1.207	0.636	1.556	1.220
TiRex-2	15	0.391	1.073	0.285	−0.044	0.725
Chronos-2 synthetic-only	13	0.401	0.297	0.125	−0.022	0.197
Chronos-2	24	0.235	0.204	0.029	−0.035	0.101
Moirai-2.0-R-small	13	2.302	−0.217	−0.735	−0.351	−0.457

Comparison scope. Window counts differ: TabPFN-TS rests on 400 windows (seeds 42, 43), TiRex-2 on 200, and TabPFN on 50 windows (seed 0) at L=256, H=48, a different window geometry; the other readers on 400 or 900. Under the shared-cell score TabICL (1.220) and TabPFN (1.137) swap places, with TabPFN-TS present or absent.

C.3 Implementation, variants and provenance

Remark A1 (Input-preserving abstention)**.** *Routing uses context, profile, origin and origin-available quality; horizon target and covariate values do not select transforms. With no supported action, including missing, mismatched or below-floor profiles, target context and raw covariates pass through unchanged. Applied transforms can still degrade forecasts; their effects are assessed empirically. Decision-core invariance does not require floating-point diagnostics to match byte-for-byte.*

Algorithm notation. The profile Π lists routing parameters; validity, enabled transforms and nonlinear-link attestation are additional metadata. The origin is t_0, $y^{\mathrm{ctx}} = \mathbf{x}$, and $C = [C^{\mathrm{ctx}}; C^{\mathrm{hor}}]$ stacks context and horizon rows of $\mathbf{C}$. Covariate channel indices run from 1 to K_0. For context mean μ_k^{ctx} and guarded scale s_k^{ctx}, $u_{t,k} = (C_{t,k} - \mu_k^{\mathrm{ctx}})/s_k^{\mathrm{ctx}}$ defines standardized covariates; $\mathbf{z} = N(\mathbf{x})$ denotes normalized target history. A is the ordered selection of basis functions; $A(C)$ evaluates them on context and horizon, and $[C, A(C)]$ concatenates channels. The screen residual is $r = R_{\mathrm{nat}}(y^{\mathrm{det}})$. $\hat{\ell}_k$ is the context-SSE-minimizing lag over permitted links and lags 0–64; $\mathrm{ACF}_k(j)$ is the mean-

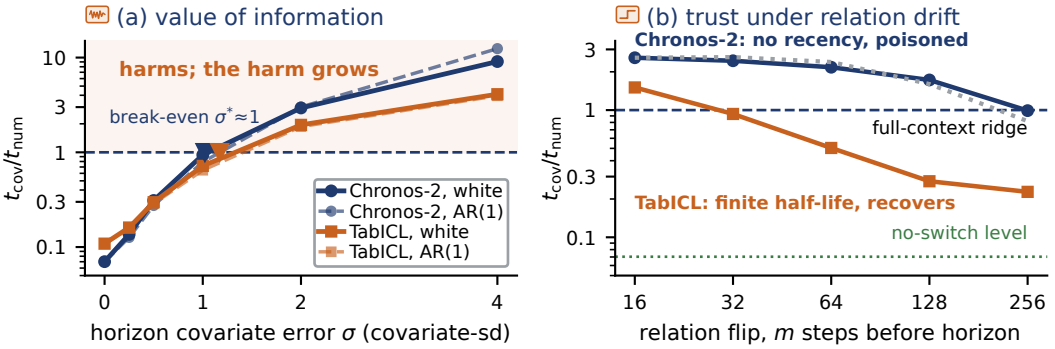

Figure 4: **Value of information and trust dynamics** (charts for the §4 prose). **(a)** $t_{\text{cov}}/t_{\text{num}}$ vs. horizon covariate error σ on both architectures: break-even σ^* (markers) sits near one covariate-sd, and past it harm *grows*. **(b)** A relation *flip* m steps before the horizon: Chronos-2 stays poisoned until the new regime fills the context; TabICL's best tested ridge fit has a $\sim$64-step half-life. Dotted, the full-context ridge; green dotted, the no-switch level.

Table 9: **Execution contract.** Numerical transform rules are in Algorithm 1. Offline quality requires calibration end $\leq$ observation time $\leq$ origin.

Stage	Required behavior
Route	Per window, truncate both contexts, realign permitted lags, shrink, then expand; invalid/inactive profiles return raw; v5 requires at least one attested native nonlinear link.
Freeze	Record and hash decision cores before forecasting; keep floating-point diagnostics separate. Bucket by action signature and retained shape.
Apply	Forecast retained target context and compiled covariates with frozen weights; reassemble in origin order.
Audit	Preserve decision-core hashes under horizon-covariate and later-window perturbations, frozen replay, and single-window/batch routing. Horizon targets are absent from the router API; classical comparators never enter selection.

centered sample autocorrelation of retained channel context at lag j. The noise scale σ_k is provider verification error in covariate-standard-deviation units; $\sigma^\star$ is the profiled break-even threshold.

Transform definitions. Algorithm 1 uses context targets and covariates, the frozen profile, and origin-available quality to select transforms. Standardize each covariate by its context mean and scale, then append products, squares and absolute deviations in the fixed permitted order. Re-center added channels; the total channel cap is 12 (2 for the conservative profile). Residualize both candidate and OLS-detrended target on native channels; normalize residual covariance by total detrended target energy. The 0.10 support threshold prevents a near-perfect native fit manufacturing support from a small residual. Optional Fourier bases use the same screen. Under independent additive covariate error of variance σ_k^2, shrink only the horizon: $C_k^{\text{hor}} \leftarrow \mu_k^{\text{ctx}} + (C_k^{\text{hor}} - \mu_k^{\text{ctx}})/(1+\sigma_k^2)$, using offline provider verification error. For truncation, $\hat{\beta}_{[a,b)}$ is the vector of separate per-channel OLS slopes for the detrended target on segment $[a, b)$. Maximize $\|\hat{\beta}_{[0,s)} - \hat{\beta}_{[s,L)}\|_2$ on a 16-step grid; require the 0.99 quantile of 32 circular context shifts (drawn once, shared across windows), a statistic at least 1.5 times that threshold, and at least 16 steps retained. Lag realignment scans $\ell \in \{0, \ldots, 64\}$ for minimum in-context SSE, requires $> 10\%$ gain, native lag budget ≤ 4 and the daily aperiodicity check in Algorithm 1. Chronos-2's budget is 24, TabICL's 4.

Method versions. Unconditional realignment caused $3.1\times$ harm on fired development windows, falsifying v2. Rounds 1–3 use v4 without the capacity gate; Round 4 uses v5. Only expansion

activated in confirmed cells. Version v6 replaces the profiled lag budget (the source of all routing divergences under swapped profiles, $8/18$ cells) with a reader-agnostic split-context guard (App. D).

Variants and mechanism limits. Replacing covariates with an in-window least-squares projection over the same dictionary improves all four synthetic deficit cells against screened expansion (0.54–0.93, ratio of median skills), but loses on 333 spring windows under corrected (1.09 $[1.01, 1.17]$) and legacy feeds (1.13 $[1.04, 1.22]$). Its 4% $[0.4, 10]$ gain on 210 summer windows is season-specific; we claim no reader-specific compilation benefit. Screened Fourier expansion repairs the oscillatory link ($1.011 \to 0.156$; synthetic-only checkpoint 0.141). On the native-linear control its ratio is 0.986 versus 1.625 for blind expansion; intervals permit smaller losses. These comparisons measure input-response behavior; they establish no internal circuit mechanism.

Protocol deviations and audit scope. Development replay used v3, Rounds 1–3 v4, and Round 4 v5. Three EIA target correlations were exposed after gain/panel predictions were committed but before manifest lock; no sign, threshold, cell or endpoint changed. One bikeshare×Chronos-2 interim ratio was exposed after decisions were frozen; that family returned a null. An output-overwrite defect detected before metric inspection required rerunning 15 bikeshare/weather cells; the 24 EIA cells were unaffected.

Round 4's initial profile/router mismatch failed validation on sixteen windows. Before evaluation, we re-locked the manifest with verified v5 profiles and router. All readers exceeded the capacity floor ($5/5$, $5/5$, $6/6$ nonlinear links); above-floor decisions matched v4 ($15/15$). A post-lock smoke-test summary was exposed and excluded from scoring; decisions remained unchanged. The dictionary and context-sweep follow-ups used earlier study outcomes to inform their designs, then fixed their own predictions before testing; the context-sweep held-out forecasts followed that lock. Independent reconstruction of the TiRex-2/Moirai-2 results verified all record seals, input hashes and prediction hashes and reproduced every MSE exactly.

Diagnostic validation. The `covprobe/` CLI implements the covariate-reader contract of §3 with registered generators. Across 25 checkable cells, reference skills agree within 8×10^{-14}; the seven-reader leaderboard regenerates bit-exactly. A historical live Chronos-2 check on two links at 40 windows reproduces reference skills exactly and the surplus signature.

Evidence tiers. Confirmed hypotheses pass the preregistered effect test (including specified Holm correction), CI and all audit conjuncts; withheld hypotheses pass the effect test but fail an audit. Not confirmed means the registered effect bar fails. Development is exploratory; descriptive covers secondary/off-primary summaries. Controlled synthetic tests support the mechanism, ablations support attribution, and prediction agreement means a frozen prediction agreed with the outcome without a registered effect hypothesis. Community-benchmark results retain the benchmark's endpoints. These tiers do not confer real-data confirmation. A CI-separated gain has upper bound < 1; confirmed harm has lower bound > 1.05. A lower bound in $(1, 1.05]$ identifies degradation but bounds no loss at 5%.

D Context dependence and transfer

Paired forecast-vintage and seasonal design. We froze source, inputs, profile, checkpoint and origin hashes before the 8 September 2026 study. The 56 cells cross fourteen Round-2/3 authorities, spring/summer and legacy/corrected feeds, with Chronos-2, L=512, H=96 and policy v5. A fresh EIA snapshot retains $333/338$ spring origins (five SCEG windows lack target hours); summer retains all 210 origins, fifteen per authority from 1 July through 26 August, with horizons ending by 29 August. This is temporal transfer on previously studied regions. Both feeds share targets, target context, day-1 weather context, UTC-hour calendar channels and context-only scaling. Missing required hours exclude both paired windows; no values are filled or timelines compressed. The updated demand snapshot and calendar make spring a new paired comparison, not an exact archive replay.

Availability assumption. For horizon offset $h \in \{0, \dots, 95\}$ the corrected feed uses $d(h) = \max\{1, \lceil (h+6)/24 \rceil\}$, so lead days 1–5 supply $19, 24, 24, 24, 5$ hours. Its forecast reference is

Table 10: **Complete paired vintage and seasonal evaluation.** Chronos-2, 14 regions per row, $L512/H96$. Entries are equal-region geometric MSE ratios with 95% hierarchical region/origin-block bootstrap intervals (4,000 resamples; seed 20260908); lower is better. The estimator averages per-origin log ratios within each region before averaging regions. Static denotes the fixed expansion bundle. Only the daggered CovLift/raw entry is the registered primary; every other entry is descriptive. The periods share regions, and the feeds share origins within each period; neither supplies additional independent regional replications.

period	weather feed	origins	CovLift/raw	CovLift/static
spring	legacy day 1	333	$0.8395\ [0.7570, 0.9164]$	$0.9481\ [0.9076, 0.9890]$
spring	corrected	333	$0.8615\ [0.7970, 0.9273]$	$0.9666\ [0.9378, 0.9947]$
summer	legacy day 1	210	$0.9859\ [0.9467, 1.0241]$	$0.9448\ [0.8576, 1.0267]$
summer	corrected	210	$0.9685\ [0.9224, 1.0065]^{\dagger}$	$1.0197\ [0.9430, 1.0958]$

$t + h - 24d(h)$ and assumed availability six hours later is no later than origin t. Provider publication timestamps remain unobserved: this removes future reference times under the registered delay assumption, without verifying release latency or reconstructing a single forecast run.

All 56 cells pass registered reference-time and provenance checks. The summer corrected-feed primary fails (p=0.05049 against 0.05); other contrasts are descriptive. Corrected/legacy absolute error ratios are $1.2571\ [1.1530, 1.3750]$ in spring and $1.5812\ [1.3733, 1.8594]$ in summer. The normalization floor is the numeric one-liner floor, shared within feed pairs, not the checkpoint's numeric-only error. All thirteen CI-separated improvement cells pass the frozen-routing shuffle criterion (per-cell median-normalized shuffled/compiled ratio > 1.05). No cell has harm CI lower > 1.05, but summer corrected-feed SRP is $1.0697\ [1.0007, 1.1489]$; this is no 5% noninferiority guarantee. No new demand-forecast tolerance pool was tested, so this study supplies no new composite confirmation verdict.

What the profile comparison identifies. A generic control retains the ordered dictionary, residualized screen, support threshold, incremental margin and channel cap, but omits reader-specific permissions and other transforms. It gives $0/1{,}086$ differences in actual reader inputs over two feeds and 543 distinct region–origin windows; decision differences are metadata only. Expansion is the only executed transform ($303/333$ spring and $89/210$ summer origins per feed). FPC, SCEG, SOCO and TVA abstain on all fifteen summer origins, with exactly equal raw and compiled error arrays. These gains do not identify incremental profiling value beyond that matched screen, nor the dictionary's design value; the held-out diagnostic test addresses a separate claim.

D.1 Context-sweep design and reference

The certified spring population contains 328 windows; the L=4,096-eligible subset contains 299 windows in thirteen regions. Here K counts appended channels, so total width is $K_0 + K$ with $K_0 = 3$. Write $r(K, L)$ for the paired geometric engineered/raw MSE ratio at context length L. The dictionary has 27 candidates (products, squares, cubes, quartics, absolute deviations and Fourier terms of three native channels). Top-K appends the highest-support channels, reselected at each L; fixed-K retains identities chosen at L=512. Results use paired-window geometric MSE ratios and moving-block bootstrap intervals (block eight, 2,000 draws). Registrations and independent reconstruction preceded interpretation; fitted forms are descriptive unless tested on held-out populations.

The stylized variance contrast for three native and K appended channels is

$$\log \frac{1 + (3 + K)/128}{1 + 3/128} - \log \frac{1 + (3 + K)/1{,}024}{1 + 3/1{,}024}.$$

It uses the $1 + p/L$ approximation for a correctly specified homoscedastic linear model, effectively independent observations and matched context/horizon second moments, as in a whitened orthogonal design. It is neither a matched-reader measurement nor a lower bound for these nonlinear, dependent data. Tables 11 and 12 give the contrasts and underlying budget curves; Tables 13–14 retain the reader, population and comparator distinctions behind the central finding.

Table 11: **Context dependence against the stylized variance benchmark.** Observed $\log r(K, 128) - \log r(K, 1{,}024)$ over the stylized variance benchmark for three native and K appended channels, spring certified population (328 windows), 95% moving-block intervals; top-K re-selects channels at each L, fixed-K keeps the identities chosen at L=512. The contrast was fixed before every held-out result.

Arm	K	observed (log), day-1 feed	stylized variance benchmark
top-K	2	0.218 [0.146, 0.300]	0.013
top-K	4	0.297 [0.214, 0.387]	0.026
top-K	8	0.388 [0.303, 0.478]	0.052
top-K	16	0.335 [0.241, 0.430]	0.100
top-K	27	0.274 [0.179, 0.373]	0.161
fixed-K	2	0.165 [0.107, 0.230]	0.013
fixed-K	8	0.314 [0.229, 0.407]	0.052

Arm	K	ratio, day-1 feed	ratio, corrected feed
top-K	2	16.5 [11.0, 22.7]	11.1 [5.7, 17.0]
top-K	4	11.4 [8.2, 14.8]	7.5 [4.2, 11.0]
top-K	8	7.5 [5.9, 9.3]	5.4 [3.6, 7.3]
top-K	16	3.4 [2.4, 4.3]	2.6 [1.7, 3.5]
top-K	27	1.7 [1.1, 2.3]	1.4 [0.8, 2.0]
fixed-K	2	12.5 [8.1, 17.4]	8.8 [5.0, 13.0]
fixed-K	8	6.1 [4.4, 7.9]	4.9 [3.2, 6.7]

Table 12: Channel budgets versus context on the certified spring population (328 windows, legacy feed). Each entry is the geometric paired engineered/raw MSE ratio at the same L, with its 95% moving-block interval (block eight, 2,000 draws). Top-K reselects identities at each context; fixed-eight uses the identities from $L = 512$. These budget-specific arms differ from the deployed screen in Table 13.

Arm	$L = 128$	$L = 256$	$L = 512$	$L = 1{,}024$
Top two	0.956 [0.900, 1.009]	0.921 [0.862, 0.975]	0.875 [0.813, 0.933]	0.769 [0.700, 0.838]
Top eight	1.120 [1.026, 1.218]	1.010 [0.924, 1.096]	0.920 [0.844, 0.994]	0.760 [0.688, 0.832]
Top 27	1.068 [0.981, 1.162]	1.045 [0.956, 1.142]	0.997 [0.912, 1.086]	0.812 [0.730, 0.897]
Fixed eight	1.069 [0.978, 1.164]	1.015 [0.932, 1.096]	0.920 [0.844, 0.994]	0.780 [0.707, 0.856]

Fixed deployment variants. A raw/expansion 50/50 blend gives 0.957 [0.924, 0.986] on 258 windows and a stricter support gate 0.939 [0.901, 0.976] on 894 held-out windows. Per-window gain rises with added-channel support (Spearman −0.19 Chronos-2, −0.15 TabICL; abstentions account for the association; intervals exclude 0).

Known-link control and unsuccessful selectors. A registered synthetic control uses three native channels, one true square and up to fifteen independent noise channels (300 windows/cell, L=128–2,048). On Chronos-2, noise-channel cost grows with K and shrinks with L (interaction 0.325 [0.275, 0.381] for K=16 against 1); dilution exceeds the variance reference by 3.2× [2.7, 3.8], while the null world's analogue includes one. A third prediction fails: true-basis/raw error rises with L (contrast −0.105 [−0.151, −0.053]) because raw gains more near the oracle floor. TabICL pays no noise-channel cost at any tested L; leave-one-L-out favors $\log L$ on Chronos-2 without discriminating a saturating form on real data.

Table 13: **The deployed screened expansion against raw at the same context length.** Geometric-mean MSE ratio, screened over raw, paired per window at each L; below 1 favors the expansion; intervals are 95% moving-block bootstrap intervals (block 8, 2000 draws, seed 17). The spring rows use the certified population (14 EIA regions, 328 windows); the long-context subset is the 299-window subpopulation eligible at L=4096 (SCEG and eight origins each of CPLE and FPC drop out), which continues to 0.630 [0.558, 0.706] at L=2048 and 0.619 [0.550, 0.692] at L=4096. Round 4 (16 regions, legacy feed), winter and summer (causal feed) are the held-out populations. The summer row prints the equal-region hierarchical bootstrap intervals of the registered seasonal estimator (regions resampled with replacement, moving origin blocks within regions; 4000 resamples, seed 20260912; Table 10); on this balanced design its point estimates coincide with the window-level ones. TiRex-2 and Moirai-2.0-R-small are computed from their per-window cells with the same estimator. Figure 3a plots the spring rows.

reader, population, n	L=128	L=512	L=1024
Chronos-2, spring, legacy, 328	0.962 [0.894, 1.032]	0.837 [0.772, 0.899]	0.727 [0.659, 0.793]
Chronos-2, spring, causal, 328	0.973 [0.908, 1.039]	0.863 [0.809, 0.923]	0.788 [0.725, 0.852]
Chronos-2, spring, legacy, long-context subset, 299	0.965 [0.891, 1.039]	0.853 [0.790, 0.919]	0.738 [0.670, 0.807]
Chronos-2, Round 4, legacy, 387	1.067 [1.012, 1.124]	0.898 [0.853, 0.943]	0.779 [0.723, 0.836]
Chronos-2, winter, causal, 202	1.074 [0.972, 1.173]	0.976 [0.915, 1.044]	0.979 [0.923, 1.036]
Chronos-2, summer, causal, 210	1.013 [0.974, 1.059]	0.969 [0.923, 1.006]	0.991 [0.958, 1.022]
TabICL, spring, legacy, 328	0.967 [0.938, 0.993]	0.922 [0.897, 0.947]	0.896 [0.866, 0.925]
TiRex-2, spring, legacy, 328	0.980 [0.946, 1.010]	0.940 [0.883, 0.995]	0.894 [0.836, 0.949]
Moirai-2.0-R-small, spring, legacy, 328	1.203 [1.150, 1.259]	1.125 [1.069, 1.189]	1.066 [1.015, 1.125]

Ten registered development studies did not establish a selector superior to the context screen. These studies tested redundancy removal, fixed rotation, retail transfer, historical calibration and nonlinear decomposition. Their outcomes range from inconclusive gains to confirmed losses against their respective controls. Across 87 cases with eight matured backtests each, a three-action hindsight oracle beats rolling validation by 8.6% (0.914 [0.869, 0.955]); local synthetic probes do not (1.011, interval includes one), nor does matured-history calibration (1.009 [0.979, 1.043]). Rolling-ridge and TabICL controls likewise establish no additional selector or expansion benefit. With official day-ahead covariates on 52 Great Britain origins, expansion equals raw on 50 dates (1.005 [1.000, 1.015]).

Reader ordering and seasonal scope. On the same 338 spring windows, equal-region geometric screened/raw ratios follow the ordering predicted before the TiRex-2/Moirai runs: Chronos-2 0.82 [0.74, 0.90], TabICL 0.93 [0.90, 0.97], TiRex-2 0.95 [0.88, 1.02], Moirai-2.0 1.11 [1.05, 1.18]. Raw/numeric-only ratios (0.51, 0.75, 0.55, 1.91) misorder TiRex-2 and TabICL; the offline reference orders them correctly. Moirai's capacity floor was waived, so this is a diagnostic comparison, not a policy-v5 safety test. A post-hoc, fired-only analysis finds monotone ratios of median skills across nonlinearity terciles (0.980, 0.938, 0.803; last [0.711, 0.854], 697 records/394 windows), unlike the calendar placebo. Spring origins recur across two feeds; these are dependent records, and pooled p-values are anti-conservative. The registered unconditional endpoint is weak: ties pin the least nonlinear tercile, the highest tercile is entirely spring, and a calendar placebo reproduces its contrast. Under the legacy feed alone (no duplicate origins), the highest within-spring tercile is 0.774 [0.677, 0.868]. Winter's registered equal-region screened/raw result (210 windows, corrected feed) is 0.96 [0.86, 1.07] for Chronos-2 and 0.95 [0.91, 0.99] for TabICL. The nonlinearity-share prediction

Table 14: **The combined configuration (weather-only dictionary, $L{=}1024$) against the deployed configuration (screened, $L{=}512$), against the screened expansion at $L{=}1024$, and against raw.** Paired geometric-mean MSE ratios for Chronos-2; below 1 favors the combined configuration. Daggered cells are the registered tests at their Bonferroni level: the two spring tests at 98.75% and the held-out test at 99.545%; every other interval is a 95% moving-block interval. The held-out test is confirmed on Round 4; on winter and summer its interval includes 1. Single components: long context alone (screened 1024 over screened 512) is 0.857 $[0.818, 0.898]$ on the legacy feed; hygiene alone (weather-only dictionary at $L{=}512$ over screened 512) is 0.965 $[0.938, 0.986]$ (registered, 98.75%).

population, feed	n	combined ($L{=}1024$) over	
		deployed (512)	screened (1024)
spring, legacy	328	0.845 $[0.797, 0.899]^{\dagger}$	0.987 $[0.968, 1.004]^{\dagger}$
spring, causal	328	0.901 $[0.841, 0.966]^{\dagger}$	1.000 $[0.984, 1.018]^{\dagger}$
Round 4, legacy	387	0.841 $[0.794, 0.891]^{\dagger}$	0.980 $[0.969, 0.992]$
winter, causal	202	0.910 $[0.816, 1.029]^{\dagger}$	1.017 $[0.999, 1.036]$
summer, causal	210	0.941 $[0.871, 1.032]^{\dagger}$	1.012 $[1.004, 1.021]$

population, feed	n	combined ($L{=}1024$) over	
		raw (512)	raw (1024)
spring, legacy	328	0.707 $[0.642, 0.769]$	0.717 $[0.649, 0.783]$
spring, causal	328	0.778 $[0.717, 0.839]$	0.788 $[0.726, 0.850]$
Round 4, legacy	387	0.755 $[0.708, 0.804]$	0.764 $[0.711, 0.817]$
winter, causal	202	0.888 $[0.824, 0.966]$	0.996 $[0.937, 1.056]$
summer, causal	210	0.912 $[0.854, 0.978]$	1.003 $[0.980, 1.025]$

fails there (median relation R^2 0.21 versus 0.51 in spring). Within-season association does not give a transportable effect law.

Additional scope checks. At $H{=}24$, all four readers use day-ahead load/wind covariates on two electricity-price markets (raw/numeric-only 0.38–0.91), with no reader gaining on both. On three ProEnFo load datasets, three readers leave expansion at 1.00–1.02. A registered reader check forecasts the final $2H$ context steps under raw and compiled inputs and keeps compilation only with a 10% backtest win. On seven held-out fev tasks, expansion loses 5.6% $[1.4, 10.5]$ against raw; the check loses 0.1% $[-0.4, 0.6]$. Across all 23 tasks it changes -2.2% to $+0.2\%$ $[-0.1, +0.7]$ (two tasks use 100-series subsamples, accounting for the original record's -2.1%). Costs against expansion are 1.025 $[0.994, 1.064]$ in summer and 1.020 $[0.950, 1.095]$ in winter: the registered outcome is partial.

Reader-agnostic lag guard. Version v6 fits the lag on the first two thirds of retained context, requires at least 10% residual-error improvement from realignment on the held-out last third, and requires the lag re-estimated there to agree within ± 2 steps. It reads context only, with no reader-specific field. On the registered controlled replay it opens on 91–94% of true-lag windows and none of $4{,}500$ no-true-lag windows, but exceeds the registered 10% per-cell-oracle tolerance on all three Chronos-2 true-lag cells and two of three TabICL cells (excesses 12–21% and 6–18%). Thus its structural screen separates these cases without meeting the registered accuracy target.

Measured basis-expansion association. The 63 cells cross 21 relations with three unscreened basis orders (2, 3, Fourier), using Chronos-2, 200 windows per cell, seed 42 and $L512/H96$. The gain is $g = \log(q_{\mathrm{raw}}/q_k)$ for FM-own ratio q; the registered predictor $c = R^2_{\mathrm{lin}}(k) - R^2_{\mathrm{lin}}(1)$ is the difference between median context fits. The registered rank test passes (Spearman 0.927, $p = 1.3 \times 10^{-27}$), but the fitted slope 2.06 $[1.69, 2.38]$ misses the registered $[0.7, 1.3]$ band. Re-expressing the predictor in matching log-variance units, $c' = \log[(1 - R^2_{\mathrm{lin}}(1))/(1 - R^2_{\mathrm{lin}}(k))]$, gives $g = 0.70c' + 0.06$, slope

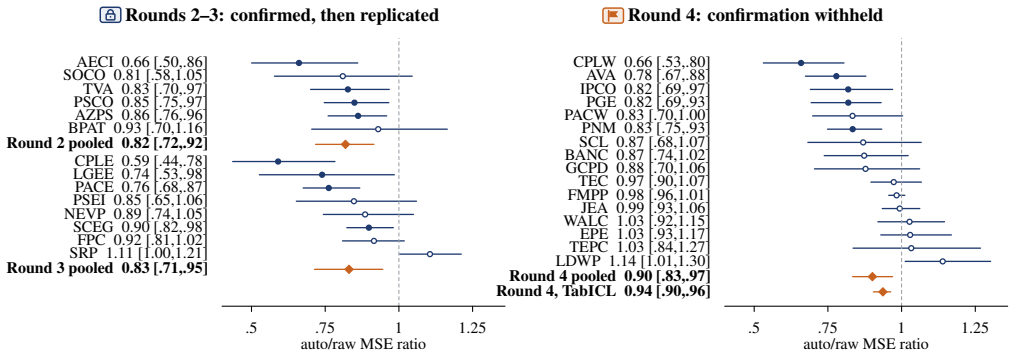

Figure 5: **Regional error ratios, Rounds 2–4.** Chronos-2 load forecasts at $H{=}96$: auto/raw MSE ratios with hierarchical-bootstrap CIs, sorted within round. Filled points have CIs below one; diamonds show pooled estimates (Table 1). Round 4 confirmation was withheld because the shuffle criteria were not met.

interval $[0.59, 0.80]$ and Spearman 0.914. This second fit is post-hoc. The association motivates expansion but establishes no identity law or guarantee that a good context fit transfers to the horizon.

E Evaluation controls and limits

Estimands and scope. The controlled tests report ratios of median normalized errors, each relative to that reader's numeric-only baseline. Development replay and stress tests report ratios of median skills with moving-block 95% intervals. Prospective evaluation and routing ablations instead report geometric MSE ratios with hierarchical-bootstrap 95% intervals. These estimators cannot be interchanged: capital bikeshare $\times$ Chronos-2 gives 1.184 versus 0.998 on the same $H{=}24$ arrays. We consolidate secondary per-cell catalogs here into component results, registered decisions and boundary cases; the primary hypotheses remain in Table 1 and all Round-2–4 primary regional effects are shown in Fig. 5.

Prospective population and seasonal split. Rounds 2–4 use the last 30% of a fixed 1 August 2025–30 June 2026 range, giving 17 March–29 June 2026 test origins. Round 2 fixes six regions and Round 3 the next eight; Round 4 fixes sixteen further authorities. Every decision is made from its context before forecasting, and paired arms use the same non-overlapping origins. Geography is replicated, but the rounds share a shoulder-season calendar: in that range the temperature–load relation bends around its minimum. The same-date spatial reanalysis in Table 18 does not create a seasonal replication. The legacy feed uses the day-1 forecast throughout context and horizon; at $H = 96$ roughly three quarters of horizon values therefore carry forecast reference times after the origin. The registered results retain this feed qualification. The separately locked paired-feed study in App. D addresses reference-time availability and summer transfer; its unobserved provider publication times remain a limitation.

Tolerance and attribution checks. Pooled tolerance covers the registered tolerance-role families and cells not predicted to gain. Rounds 1–4 yield compiled/raw geometric MSE ratios with 95% hierarchical-bootstrap CIs of 0.985 $[0.959, 1.004]$, 0.989 $[0.974, 1.001]$, 0.981 $[0.964, 0.998]$, and 1.001 $[0.996, 1.007]$. All upper bounds meet the registered ≤ 1.02 threshold.Tolerance is an aggregate endpoint, not a bound on every region or transformed window. All comparisons include abstained windows, and unchanged input is distinguished from a statistical non-degradation claim. Decision-invariance audits freeze routing before covariate derangement; the deranged forecast then tests reliance on horizon covariate content. The same-input classical panel comprises ridge, GBM, sparse LassoCV and hierarchical empirical Bayes; panel superiority requires beating all four. Pilot rows do not enter the reported confirmatory counts. The per-cell derangement criterion remains binding even when a pooled effect is significant: it is the reason both Round-4 reader claims are withheld.

Table 15: Controlled repairs: raw → compiled ratios relative to each checkpoint's numeric-only error (300 windows per condition). Shrinkage entries are each arm's separate maximum over nine noise conditions, using the true noise level; they are not a paired worst-case cell.

Condition	Chronos-2	C2-synth	TabICL
Interaction	$0.965 \to 0.169$	$0.845 \to 0.111$	$0.479 \to 0.105$
Shrinkage	$11.109 \to 0.794$	$9.763 \to 0.651$	$4.093 \to 0.773$
Flip, m=16	$2.793 \to 0.191$	$2.215 \to 0.127$	$1.386 \to 0.409$
Flip, m=32	$2.395 \to 0.137$	$2.081 \to 0.109$	$0.899 \to 0.321$
Flip, m=64	$2.161 \to 0.139$	$1.894 \to 0.104$	$0.561 \to 0.184$
Lag 48	$1.107 \to 0.063$	$0.983 \to 0.088$	$0.996 \to 0.106$
Lag 64	$1.123 \to 0.083$	$0.990 \to 0.098$	$0.997 \to 0.120$

What repair establishes. At lag 48, compiled moving-block intervals are $[.054, .079]$, $[.066, .109]$ and $[.090, .127]$ for Chronos-2, C2-synth and TabICL. Classical comparators receive the same covariate information, but different transformations: GBM uses raw covariates, ridge the same truncated context, and lag-scanning OLS raw covariates. Ridge has lower point error at m=64 ($0.122/0.094/0.116$ versus compiled $0.139/0.104/0.184$). Shrinkage requires calibrated verification error: using $\sigma/2$ or 2σ yields $1.059/0.814$ on Chronos-2, $0.907/0.684$ on C2-synth and $0.960/0.810$ on TabICL. Thus the component repair does not establish dominance over classical methods or robustness to noise miscalibration.

Table 16: Stress-test limits. The four traps have real context relationships but uninformative, inverted or spiked horizons. Upper-CI thresholds are 1.05 per family and 1.02 pooled. The table retains all per-family exceedances and the matched useful-covariate control.

Reader	Exceeding family: CI upper	Useful control ratio [CI]
Chronos-2	none	0.399 $[0.349, 0.431]$
C2-synth	heavy tail: 1.068	0.393 $[0.345, 0.438]$
TabICL	none	0.920 $[0.816, 0.966]$
Moirai-1.1	coupling: 1.073; tail: 1.084	1.010 $[0.967, 1.061]$

The router abstains on 96–99% of trap windows, with matching rates across readers. The traps are linear in context: this suite does not test a nonlinear relationship that breaks only in the horizon, and establishes no spuriousness-detection ability. These stress results precede v5's capacity gate, which returns raw inputs on every tested Moirai-1.1 window; input-preserving abstention is distinct from statistical safety.

Table 17: Development replay: falsification and revised boundaries. Ratios of median skills with moving-block 95% intervals; these are development results, not prospective confirmations.

Policy	Cell / reader	Auto/raw [CI]
v2	CISO load / Chronos-2	1.191 $[1.006, 1.345]$
v2	CISO day-ahead / Chronos-2	1.636 $[1.251, 2.192]$
v2	CISO day-ahead / TabICL	1.617 $[1.357, 1.922]$
v2	EPF / Chronos-2	1.169 $[1.073, 1.278]$
v3	GEFCom / Chronos-2	0.845 $[0.760, 0.993]$
v3	EIA-PJM / TabICL	0.887 $[0.786, 0.999]$
v3	NYIS day-ahead / C2-synth	0.970 $[0.846, 0.999]$
v3	EPF / Chronos-2	1.096 $[1.022, 1.188]$

Unconditional realignment failed on periodic aliases: CISO day-ahead Chronos-2's ratio was 3.12 on realigned windows versus 0.896 on expansion-only windows. A label-keyed exemption had hidden this failure in v1; we retired that design, registered expansion-primary v3 and reran all 39 development cells. Revised CISO day-ahead ratios are $1.00/0.94/1.03$ (Chronos-2/TabICL/C2-synth). All 39 decisions pass invariance; all three CI-separated lifts collapse under derangement. GEFCom and PJM

also beat the same-input classical panel. NYIS's lift, on a natively linear channel predicted to abstain (72% abstention), supplies no confirmatory efficacy evidence. EPF remains a disclosed, preregistered unfixed boundary.

Table 18: Prospective boundary cases and descriptive spatial robustness. Cell estimates are geometric auto/raw MSE ratios with 95% intervals. CI lower > 1.05 defines confirmed harm. The last three rows pool the fourteen Round-2/3 load regions; their intervals retain common dates.

Case	Chronos-2	TabICL
Round 1: capital bikeshare	0.998 [.953, 1.044]	1.062 [1.005, 1.123]
Round 1: citi bikeshare	1.096 [1.032, 1.172]	1.010 [.946, 1.088]
Round 3: SRP load	1.105 [1.003, 1.212]	—
Round 4: LDWP load	1.139 [1.012, 1.302]	1.003 [.971, 1.035]
Round 4: WALC load	1.027 [.919, 1.146]	0.933 [.898, .978]
Round 2/3: geographic clusters	0.821 [.743, .896]	0.943 [.905, .984]
Worst leave-one-region-out	0.847 [.772, .920]	0.943 [.908, .978]
WECC interconnection only	0.886 [.802, .981]	0.941 [.895, .988]

Round 4 scores all 96 cells on sixteen new authorities. Both pooled primary tests pass Holm correction (m=2), but WALC $\times$ TabICL's deranged/auto ratio 0.998 fails the shared $>$1.05 causal-attribution requirement. Both confirmations are withheld; the confirmed reader count remains one. None of the listed adverse cells establishes confirmed harm. The pooled tolerance ratio is 1.001. The geographic reanalysis groups same-metro/adjacent pairs into nine clusters; the worst exclusion drops CPLE. Eastern-only intervals also exclude one (0.769 [.659, .876] Chronos-2; 0.928 [.878, .981] TabICL). These descriptive checks address spatial dependence, not seasonal transfer or the legacy feed's vintage issue.

Table 19: Routing ablation: unconditional expansion bundle versus CovLift-Auto. Entries are bundle/router geometric MSE ratios [hierarchical-bootstrap 95% CI]; > 1 favors routing. This retrospective ablation is distinct from the registered hand-built practitioner baseline, which remains unrun.

Stratum	Chronos-2	TabICL
Rounds 2–3 load	1.059 [1.013, 1.106]	0.973 [.944, 1.002]
Round 1 load	1.011 [.968, 1.060]	0.982 [.940, 1.015]
Round 1 day-ahead	1.012 [.954, 1.070]	0.949 [.907, .986]
Round 1 bikeshare	0.913 [.852, .969]	0.969 [.919, 1.031]
Round 1 weather	1.147 [1.099, 1.201]	1.125 [1.059, 1.194]

The router is no worse than the bundle on $12/14$ Round-2/3 Chronos-2 load cells but only $5/14$ TabICL cells. Weather has no predicted headroom; routing avoids the unconditional bundle's degradation. The separate classical-panel comparison finds wins on all 30 Chronos-2 load cells in Rounds 2–4. This does not isolate profile-informed selection: the matched generic dictionary/screen/cap control gives identical inputs at all 543 origins.

Retail selection test. The registered Rossmann experiment retains 27 of 32 hash-ordered stores; five lack Sales. Each has twelve fortnightly origins from 14 February to 18 July 2015, L=512, H=14; the first four provide initial historical validation, and the last eight dates (11 April–18 July) give 216 outer decisions. Eligibility requires finite Sales, Promo and DayOfWeek; closed-day zeros remain. Promo availability is assumed without announcement vintages; Customers and Open are omitted, closures do not zero forecasts, and pretraining exposure is unknown.

The H=96 utility coefficients and thresholds transfer without retail refitting. The structured readout predicts action utility using the measured reader profile; the generic no-profile control uses context descriptors without that profile. Calibration uses 384 native calls across six Hermite degrees, sixteen independent worlds per degree and autocorrelation settings 0.3 and 0.9, frozen before retail execution. Each degree averages $\log\{\max(\mathrm{MSE}_{\mathrm{raw}}, 10^{-12})/\max(\mathrm{MSE}_{\mathrm{num}}, 10^{-12})\}$. The current context's

Table 20: All four retail co-primary contrasts: structured selector/comparator geometric paired error ratios. Equal weights across 27 stores and eight dates; two-sided 98.75% intervals implement the registered Bonferroni nominal 95% familywise convention.

Comparator	Ratio	Interval
Numeric-only	0.812626	$[0.705371, 0.927633]$
Fixed screened expansion	0.997628	$[0.964056, 1.028020]$
Four-fold rolling selection	1.011983	$[0.979283, 1.041809]$
Generic no-profile selector	1.001956	$[0.996003, 1.009764]$

mean signed lag-one covariate autocorrelation chooses the nearer setting, with ties to 0.3; this rule does not assume an AR data-generating process for retail covariates.

The eight selectable actions are raw, fixed screened expansion, low- and high-order polynomials, two Fourier dictionaries, interactions and splines. Transformations retain both raw channels and use context-only screening, an eight-addition cap and fixed clipping; numeric-only is a ninth comparator outside the selectable library. Fixed screened expansion is the dictionary control, not the full CovLift-v5 router. All selectors reuse the same action forecasts. Rolling selection minimizes the preceding four matured folds' mean log error relative to raw, with 10^{-12} floors; learned selectors retain their frozen zero-utility threshold. Every selector projects to the first representative of each exact current-input equivalence class, with raw-first ties. Choices are sealed before outer outcomes; earlier outer horizons can enter history only once matured. All nine arms are evaluated at all twelve origins (2,916 retail forecasts). Each paired loss is $\log(\mathrm{MSE} + 10^{-12}\max\{\mathrm{Var}(y_{\mathrm{context}}), 1\})$. The bootstrap independently resamples stores and four noncircular two-date blocks from the seven possible starts, sharing the resulting eight dates across policies and stores (10,000 draws, seed 20260921); boundary dates receive less weight. Store-only intervals conditional on the observed dates are sensitivity results. Eight dates, residual store dependence and one chain limit generalization. The selector always chooses splines and matches generic no-profile inputs in $194/216$ cells. Statistical advancement requires all four upper bounds below one; practical advancement also requires all point ratios ≤ 0.98. Both fail. Intervals containing one establish neither equivalence nor harm. The numeric-only improvement supports the selected covariate/transformation pipeline while incremental selection value remains unestablished.

Table 21: External-baseline summaries, all relative to raw. n counts reader–series cells except the two $H{=}96$ pooled endpoints. Gain: 95% CI upper < 1; harm: lower > 1.05; other loss: lower in $(1, 1.05]$. Loss categories are disjoint and do not bound losses by 5%. R1 median-MSE/skill rows use moving-block intervals; geometric rows use hierarchical intervals. Per-cell gain is not composite confirmation. The maximum is over the row's counted units.

Method	Population and estimator	n
TATO	R1 cells, both readers, $H{=}24$; median MSE ratio	26
TFMAdapter	R1 cells, both readers, $H{=}24$; median MSE ratio	20
TATO	R2/3, Chronos-2, $H{=}96$; geometric MSE ratio	1
TFMAdapter	R2/3, Chronos-2, $H{=}96$; geometric MSE ratio	1
CovLift-Auto	same TATO cells; median skill ratio (MSE/floor)	26
CovLift-Auto	all R1–4; geometric MSE ratio	219

Method	n	CI-separated gains	Confirmed harm	Other losses	Max ratio
TATO	26	0	11	3	2.75
TFMAdapter	20	1	0	0	1.22
TATO	1	0	0	1	1.33
TFMAdapter	1	0	0	0	1.01
CovLift-Auto	26	1	0	0	1.20
CovLift-Auto	219	30	0	9	1.14

External baselines. At $H{=}96$ the fourteen-region, 338-origin TATO and TFMAdapter runs require $L{=}960$ (a recorded deviation); CovLift uses $L{=}512$ and separate per-round estimation on those origins. The respective ratios are 1.325 $[1.049, 1.733]$, 1.013 $[0.990, 1.038]$, and $0.819/0.831$: matching windows does not match context or estimator. At $H{=}24$, TATO uses median MSE while the matched CovLift row uses median baseline-normalized skill; these are not interchangeable. We disable TATO test-label re-ranking and time-axis rescaling; identity remains a candidate. TATO selects its transform using embargoed target labels, a budget the label-free router does not receive.

The trained-adapter gate requires beating frozen raw and the same-input classical panel; four of eight cells qualify. CovLift/adapter is 0.953 $[0.879, 1.036]$ on EPF/TabICL, meeting the registered CI-upper ≤ 1.05 match criterion. GEFCom/Chronos-2 (0.954 $[0.767, 1.141]$) and GEFCom/TabICL (1.172 $[0.910, 1.361]$) do not; FreshCov/TabICL favors the adapter (1.856 $[1.605, 2.106]$). Four invalid cells fail to improve on frozen raw, including the registered immune primary FreshCov/Chronos-2; they supply no head-to-head efficacy evidence. CovLift trains no per-target parameters; adapters train 25.8k–29.0k on 300–1200 windows. These are development-stage comparisons.