ONE MORE TOKEN: PATCH-BOUNDARY REVISIONS IN TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fixed-patch time-series foundation models group past observations into tokens before making a forecast. Adding one observation can create a new token and produce a sharp change in the prediction. We study this transition in Chronos-Bolt, TimesFM, and Moirai by extending the same history one observation at a time. Among the sampled local transitions, the token boundary has the largest average revision in 46 of 48 model, dataset, and horizon settings. Changing Moirai's patch size moves the revision peak to the new boundary. An attention intervention shows that allowing other tokens to read the new token contributes to the change, with additional TimesFM experiments confirming the effect on previously unused series. We then explain how these revisions affect accuracy. An exact squared-error decomposition separates the size of the revision from its direction relative to the model's existing error. On generated autoregressive series, Moirai's revision corrects that error, with an expected-loss change 2.47% of the irreducible forecasting variance lower than at nearby non-boundary transitions. Real-data results show that the loss effect varies across models and datasets. Together, these findings connect a discrete change in input representation to both forecast revisions and their accuracy.

1 INTRODUCTION

Time-series foundation models use patterns learned from many series to forecast new data. Several widely used models first divide the observed history into fixed-size groups called patches. Each patch is mapped to a vector, or token, which the forecasting network processes. This representation creates a simple question: how does the prediction change when the available history grows by one observation? With patches aligned to the most recent observation, an extra oldest observation can create an additional token while every more recent observation stays in place.

We examine this event by keeping the forecast time fixed and extending the history one observation at a time. Across Chronos-Bolt, TimesFM, and Moirai, prediction changes are largest near lengths that create a new token. Figure 1 illustrates the transition. The effect follows the patch size and remains visible under controls for input scaling and normalization. We then intervene inside the model by preventing other tokens from reading the new token through attention. Figure 3 shows that this intervention moves predictions back toward their values before the token appeared. An additional TimesFM study reproduces the effect on series that were not used in the earlier experiments.

To understand the effect on accuracy, we compare model predictions with the exact expected future of a generated autoregressive process. The change in expected squared error has two parts: the squared size of the forecast revision and its alignment with the model's previous error. A revision that points toward the expected future can improve the forecast. Moirai provides a clear example: it makes the largest revision among the tested models, and this revision reduces expected loss relative to nearby non-boundary transitions. Table 5 reports the risk changes, while Figure 4 shows the two contributing terms.

Our contributions connect the input representation, the attention mechanism, and forecast accuracy in one analysis. We measure the prediction change at individual token boundaries, identify the new token's attention contribution, and explain the resulting loss changes with an exact identity. The real-data results in Table 7 extend this analysis to measured future values and show how the same mechanism has different consequences across models and datasets.

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.

2 Related Work

Patching reduces the number of vectors needed to represent a long history. PatchTST uses this structure to improve long-term forecasting (Nie et al., 2023). TimesFM applies patching in a pretrained decoder model (Das et al., 2024), and Moirai supports several patch sizes within a shared forecasting model (Woo et al., 2024). Chronos-Bolt is the patch-based member of the Chronos forecasting family (Ansari et al., 2024). These models provide the setting for our study of the prediction change caused by adding one token.

Other work studies how to choose patch boundaries. EntroPE uses conditional entropy to place boundaries near changes in a series (Abeywickrama et al., 2025). Zucchi et al. (2026) compare adaptive and uniform patching under controlled forecasting experiments. We examine a pretrained model with a fixed patch size and measure what happens at each one-observation extension. This gives a direct connection between the patch layout and the model's response during prediction.

Context selection and inference-time methods also change the information presented to a forecaster. Retrieval-augmented forecasting uses relevant examples from other series (Han et al., 2025), while diversified inference combines predictions obtained from different inputs (Hua et al., 2026). GITCO identifies and suppresses harmful context patches in TimesFM (Pandey et al., 2026). Our experiments keep the forecast time and recent observations fixed, then measure the effect of the new token's attention access. This isolates a specific representation change whose accuracy effect can be calculated separately.

Internal interventions provide a further way to study forecasting models. time2time changes hidden states by transferring statistical properties between events (Sanyal et al., 2025). We keep the new token's projected content and position fixed and change whether other tokens can read it. This intervention identifies its contribution to the boundary revision and connects that contribution to the model's forecast error.

3 Experimental Setup and Measurements

3.1 Context length, tokens, and boundary sensitivity

Let x_t be the observed value at time t, where t is the last observed time, also called the forecast cutoff. A context of length L contains $\mathbf{x}^{(L)} = (x_{t-L+1}, \ldots, x_t)$. The forecast $\hat{\mathbf{y}}(L) = (\hat{y}_1(L), \ldots, \hat{y}_H(L))$ predicts the same future values $x_{t+1}, \ldots, x_{t+H}$ for every L. Here H is the forecast horizon, measured in observations, and h indexes a forecast step from 1 to H. Increasing L adds the next oldest observation and keeps the cutoff, future target, and model parameters fixed. We measure the size of an adjacent revision by

$$d(L) = \frac{1}{aH} \sum_{h=1}^{H} \left| \hat{y}_h(L) - \hat{y}_h(L-1) \right|. \tag{1}$$

The scale a is the mean absolute seasonal difference in the training data plus 10^{-8}. It stays fixed across context lengths and is also used for mean absolute scaled error (MASE) (Hyndman & Koehler, 2006). Appendix A gives its calculation and the seasonal period for each dataset.

A patch contains at most P observations, where P is the patch size. Let $J(L)$ be the number of input patch positions after padding, and let $o_j(L)$ count the observed values in patch position j. We call a context token active when its patch contains at least one observed value. The number of active context tokens and the boundary indicator are

$$K_{\text{active}}(L) = \sum_{j=1}^{J(L)} \mathbf{1}\{o_j(L) > 0\}, \qquad b(L) = \mathbf{1}\{K_{\text{active}}(L) > K_{\text{active}}(L-1)\}. \tag{2}$$

The indicator $\mathbf{1}$ equals one when its condition holds and zero otherwise. Padding-only positions and future prediction tokens are excluded from K_{active}. We obtain the counts from each model's input observation mask; Appendix B describes these measurements.

The leading token is the leftmost active context token. Let $f(L)$ count the observed values in its patch. For right-aligned contexts with no missing observations, $K_{\text{active}}(L) = \lceil L/P \rceil$, where the ceiling

rounds upward. Its observation count is $f(L) = 1 + ((L-1) \bmod P)$, and the remainder $L \bmod P$ gives the position within a patch cycle. For a positive integer k, write $B = kP$. The transition from B to $B+1$ then creates a new leading token with $f(B+1) = 1$. This token contains only the newly added oldest observation.

We define boundary excess sensitivity (BES) for each fixed series and cutoff, before combining results across examples. The boundary revision is $d(B+1)$, and its four nearby non-boundary revisions have length indices $\mathcal{N}_B = \{B-1, B, B+2, B+3\}$. Thus

$$\mathrm{BES}_t(B) = d_t(B+1) - \mathrm{median}_{L \in \mathcal{N}_B}\, d_t(L). \tag{3}$$

The subscript t emphasizes that all five revisions use the same cutoff and series. Each $d_t(L)$ compares forecasts at $L-1$ and L, so computing this contrast requires forecasts at all six lengths $B-2, \ldots, B+3$. We include a contrast only when the four control transitions are observed and have $b(L) = 0$. For four values, the median is the average of the middle two after sorting. Reported BES first takes the median across boundaries within a cutoff, then across cutoffs within a series, and finally across series.

3.2 Models and data

Table 1 lists the three primary model configurations. Each model produces one forecast value per future step using its own prediction output. The analysis compares changes within each model. We use eight datasets: generated first-order autoregressive (AR(1)), SeasonalAR, and sinusoidal series, together with ETTh1, ETTm1, Electricity, Traffic, and Weather. The real datasets come from the Time-Series-Library benchmark collection (Zhou et al., 2021; Wu et al., 2023). Appendix A describes the data sources, sampling intervals, data splits, and synthetic generators.

Table 1: Primary models. P is the number of observations per patch. The context cap is the maximum permitted history in our experiments.

Model	P	Point prediction	Context cap
Chronos-Bolt-small	16	Median quantile	2048
TimesFM 2.5 200M	32	Official point output	2048
Moirai 1.1 R-small	32	Distribution mean	Model default

The Chronos-Bolt AR(1) sweep evaluates every integer length $L = 192, 193, \ldots, 768$, with step size $\Delta L = 1$, at $H = 24$. The initial patch-size experiment evaluates every integer from 192 to 512. Local sweeps evaluate all six lengths $B-2, \ldots, B+3$ around $B \in \{256, 512, 1024\}$, at $H = 24$ and $H = 64$. The attention and loss studies use $H = 24$. The attention study uses 64 series from each of AR(1), Traffic, and Electricity, with eight cutoffs per series and $B \in \{256, 512\}$. A separate outcome study uses 64 Traffic and 64 Electricity series. The exact-risk study uses 128 additional AR(1) series. The additional TimesFM attention study uses those same AR(1) series and 64 further Traffic series. These later series do not overlap with the earlier studies. Appendix A lists the sample allocations by experiment.

3.3 Attention intervention

At each boundary, $\hat{\mathbf{y}}^0 = \hat{\mathbf{y}}(B)$ is the forecast before the new token appears, and $\hat{\mathbf{y}}^+ = \hat{\mathbf{y}}(B+1)$ is the forecast with the new token. We also compute $\hat{\mathbf{y}}^{\mathrm{mask}}$ at length $B+1$. This forecast uses the same projected token vectors, positions, and scaling state as $\hat{\mathbf{y}}^+$, while the attention mask prevents the new token from acting as a key or value. In attention, a query is compared with token keys to assign weights to token values. Masking removes the new token from these weighted reads. Its own query is retained, and its representation remains in the network. Chronos-Bolt masks this column in encoder attention and decoder-to-encoder attention; TimesFM masks it in every self-attention layer of both sign-averaging paths; Moirai masks it in encoder attention over its packed token sequence. We verify paired pre-attention tensors, positions, and scaling state by hash equality. For TimesFM, all 40 attention calls verify the target column in both sign-averaging paths. Appendix B gives the full pairing checks.

For any two forecast vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^H$, define $D(\mathbf{u}, \mathbf{v}) = (aH)^{-1} \sum_{h=1}^{H} |u_h - v_h|$, the same distance used in Equation 1. We report

$$\Delta_{\text{mask}} = D(\hat{\mathbf{y}}^+, \hat{\mathbf{y}}^{\text{mask}}), \qquad \Delta_{\text{back}} = D(\hat{\mathbf{y}}^+, \hat{\mathbf{y}}^0) - D(\hat{\mathbf{y}}^{\text{mask}}, \hat{\mathbf{y}}^0),$$
$$\rho = \frac{\langle \hat{\mathbf{y}}^+ - \hat{\mathbf{y}}^{\text{mask}}, \hat{\mathbf{y}}^+ - \hat{\mathbf{y}}^0 \rangle}{\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}^{\text{mask}}\|_2 \, \|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}^0\|_2}. \tag{4}$$

Here Δ_{mask} measures the effect of masking, $\Delta_{\text{back}} > 0$ means that masking moves the forecast closer to the earlier prediction, and ρ measures agreement between the two revision directions. The brackets denote the dot product, which sums componentwise products. The norm $\| \cdot \|_2$ is the square root of the sum of squared entries. Values of ρ near one indicate closely aligned directions. Appendix B gives the model-specific mask locations and treatment of revision vectors with zero norm.

3.4 Expected and observed forecast loss

For the generated first-order autoregressive process, $x_t = 0.8x_{t-1} + \epsilon_t$, the innovations ϵ_t are independent standard Gaussian variables. Expanding the recurrence gives

$$x_{t+h} = 0.8^h x_t + \sum_{j=1}^{h} 0.8^{h-j} \epsilon_{t+j}.$$

The future innovations are independent of the observed history and have zero mean. Their variances add, so the future value x_{t+h}, conditional on the latest observation x_t, has mean m_h and variance v_h:

$$m_h = 0.8^h x_t, \qquad v_h = \frac{1 - 0.8^{2h}}{1 - 0.8^2}, \qquad \bar{v}_H = \frac{1}{H} \sum_{h=1}^{H} v_h. \tag{5}$$

The vector $\mathbf{m} = (m_1, \ldots, m_H)$ is the exact conditional mean of the future trajectory. The average variance $\bar{v}_H$ is the expected mean squared error of this optimal mean forecast. It provides a fixed scale for comparing the model's excess squared error:

$$q(L) = \frac{1}{H\bar{v}_H} \sum_{h=1}^{H} (\hat{y}_h(L) - m_h)^2. \tag{6}$$

The model's full expected mean squared error is $(1+q(L))\bar{v}_H$. Its changes can therefore be computed directly from q, without sampling a future trajectory. For an oracle that knows this generating process, x_t already determines the conditional future distribution. Adding an older observation leaves that distribution unchanged, while a TSFM's representation and forecast can change. Such a revision can move an imperfect forecast toward the oracle mean. Using the same control indices $\mathcal{N}_B$ as BES, we define

$$\Delta q(L) = q(L) - q(L-1), \qquad \Delta q_{\text{local}}(B) = \Delta q(B+1) - \text{median}_{L \in \mathcal{N}_B} \Delta q(L). \tag{7}$$

The local risk change subtracts the median of the four nearby non-boundary risk changes from the natural boundary change $\Delta q(B+1)$. A negative local contrast means that the boundary risk change is lower than the median change at nearby non-boundary transitions. The attention contribution to risk is $q(\hat{\mathbf{y}}^+) - q(\hat{\mathbf{y}}^{\text{mask}})$, where q applied to a forecast vector uses the same squared-error formula. A positive value indicates higher expected loss with the token visible; a negative value indicates an improvement.

For real data, we evaluate $\ell(L) = (aH)^{-1} \sum_{h=1}^{H} |\hat{y}_h(L) - x_{t+h}|$, the MASE of the forecast. The local change in MASE applies Equation 7 with ℓ in place of q. Its sign compares the boundary error change with the median change at nearby non-boundary transitions: negative values indicate a lower change, and positive values a higher change.

3.5 Statistical summaries

Attention and loss comparisons are paired at the same series, cutoff, and boundary. We average boundary comparisons within a cutoff, cutoffs within a series, and series within a dataset. Cross-model averages give each model equal weight. Figures and loss tables report descriptive two-sided 95% confidence intervals. The additional TimesFM tests report one-sided 95% lower bounds and Holm-adjusted p-values for the six comparisons in Table 4. Their reference values are 0.001 for the two normalized distances and zero for directional agreement. These tests ask whether the measured effects exceed those reference values. Appendix C describes the resampling procedures.

4 Boundary Revisions and Their Effect on Forecast Accuracy

4.1 Prediction changes follow the token boundary

Figure 1 illustrates the one-observation extension. At $L = B$, each observed patch is full; at $L = B + 1$, the new leading token contains one value and K_{active} increases by one. Panel C uses 128 AR(1) series from the dense sweep. Figure 2 uses the separate 32-series patch-size study; Table 9 gives both sample designs. At the measured $256 \to 257$ transition, active-token counts increase from 16 to 17 for Chronos-Bolt and from 8 to 9 for TimesFM and Moirai (the latter two at $P = 32$); each new leading patch contains one observed value (Table 11).

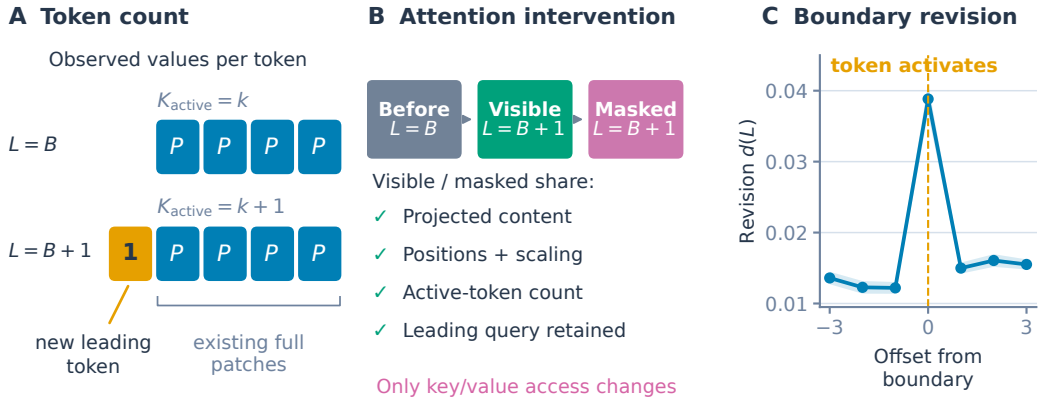

Figure 1: **Adding one observation creates a new context token.** A: the active-token count increases from k to $k + 1$. B: visible and masked inputs are paired. C: mean Chronos-Bolt revisions on AR(1); zero marks $B \to B + 1$. The band gives a 95% interval across series.

The revision peaks repeat with the patch cycle. In the local study, the boundary has the largest average revision among the five sampled transitions, $L \in \{B-1, B, B+1, B+2, B+3\}$, in 23 of 24 settings at each horizon, giving 46 of 48 overall. All 48 settings have a positive lower confidence bound for BES; Table 12 provides the counts. The two rank-two local settings are Moirai on Sinusoid at both horizons; both retain positive BES lower confidence bounds. Figure 2B shows that changing Moirai's patch size moves the peak to the corresponding new-token position. On AR(1), BES increases from 0.0063 at $P = 8$ to 0.2208 at $P = 64$; Table 3 compares the two generated datasets.

The controls in Tables 14 and 2 connect the peaks to the token layout. Seasonal-naive forecasts have maximum $d(L) = 0$ in all 16 dataset–horizon settings. For artificial periods 16 and 32, Chronos-T5's pseudo-boundary phase $r = 1$ ranks 7th and 9th on AR(1), and 13th and 22nd on Traffic, respectively. Chronos-Bolt retains its boundary peak with a fixed input scale. TimesFM retains the peak with input normalization disabled, sign averaging disabled, or both disabled. Table 2 gives the corresponding BES estimates and intervals.

Table 2: Scaling controls on 32 AR(1) series per setting. All settings retain phase rank one. Brackets are two-sided 95% BES intervals.

Model	Setting	BES [95% CI]	Rank
Chronos-Bolt	Default	0.02052 [0.01815, 0.02422]	1
Chronos-Bolt	Fixed scale	0.02098 [0.01799, 0.02329]	1
TimesFM	Default	0.01473 [0.01141, 0.01890]	1
TimesFM	No normalization	0.01473 [0.01141, 0.01885]	1
TimesFM	No sign averaging	0.01954 [0.01563, 0.02341]	1
TimesFM	Neither operation	0.01954 [0.01560, 0.02346]	1

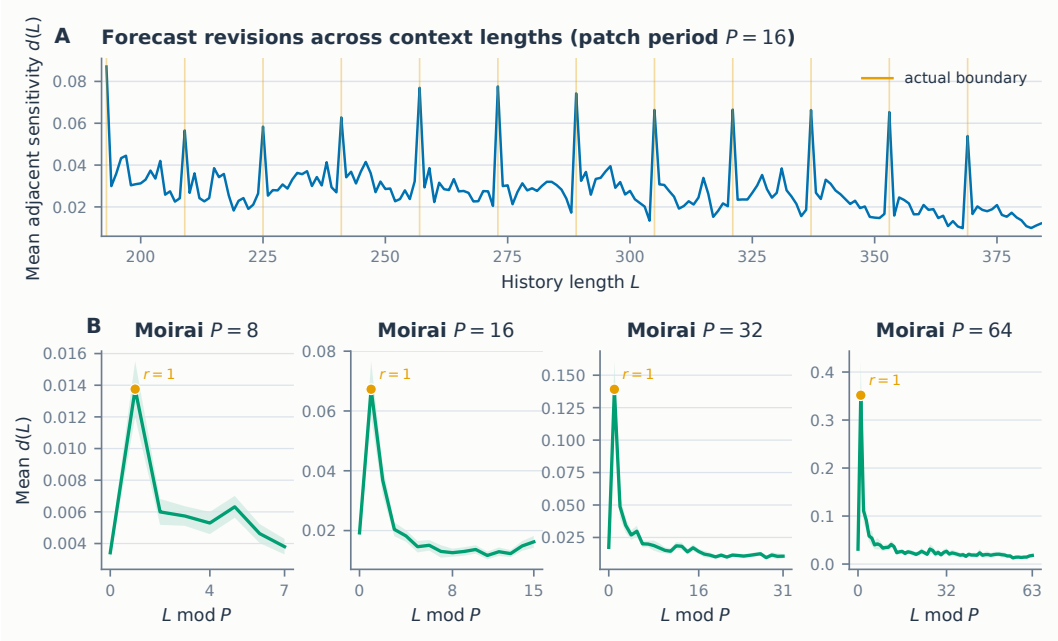

Figure 2: **Forecast revisions track the patch layout.** A: Chronos-Bolt revisions over part of a dense context sweep. B: Moirai revisions grouped by position $r = L \bmod P$ within the patch cycle. The gold marks identify the new-token position $r = 1$. Both panels use generated AR(1) series described in Appendix A. Bands in B are 95% intervals based on variation across series.

Table 3: Chronos-Bolt dense results and Moirai patch-size intervention on 32 series per dataset, with $H = 24$ and $L = 192, \ldots, 512$. BES measures local excess revision; brackets give two-sided 95% intervals from 2,000 series-bootstrap replicates. Rank covers the full patch cycle.

Model	P	AR(1)		Sinusoid	
		BES [95% CI]	Rank	BES [95% CI]	Rank
Chronos-Bolt	16	0.0214 [0.0167, 0.0276]	1	0.0152 [0.0133, 0.0188]	1
Moirai	8	0.0063 [0.0053, 0.0074]	1	0.1586 [0.1503, 0.1671]	1
Moirai	16	0.0312 [0.0272, 0.0412]	1	0.0220 [0.0045, 0.0379]	2
Moirai	32	0.0801 [0.0702, 0.0907]	1	0.0281 [0.0239, 0.0309]	1
Moirai	64	0.2208 [0.1296, 0.2921]	1	0.0854 [0.0807, 0.0911]	1

Across Moirai's four patch sizes, both generated datasets retain positive BES lower confidence bounds, while the effect magnitude varies with the generating process. This separate dense study ranks the full patch cycle; the local study above compares five neighboring transitions.

4.2 Reading the new token contributes to the revision

Figure 3 compares visible and masked predictions for all nine model–dataset combinations. Masking changes the forecast ($\Delta_{\text{mask}} > 0$) and reduces its distance from the earlier prediction ($\Delta_{\text{back}} > 0$) in every combination. The direction agreement ρ shows that the removed contribution points along the natural boundary revision. Table 13 gives the three measurements for each combination. The intervention keeps projected content, positions, scaling, and the observation-based token count fixed. This supports a causal contribution from the new token's key/value visibility. Figure 3 uses 5,000 hierarchical bootstrap replicates, resampling series and then cutoffs within each selected series while retaining paired conditions and boundary comparisons together.

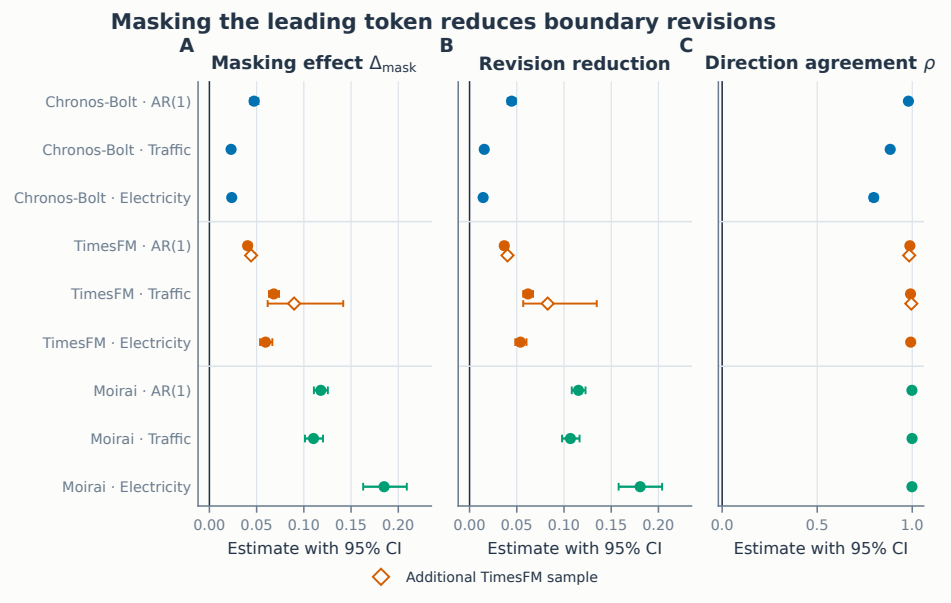

Figure 3: **Masking the leading token reduces boundary revisions.** Circles: original estimates with two-sided 95% confidence intervals. Diamonds: additional TimesFM estimates on previously unused series. Panels use Equation 4; Tables 13 and 4 give the full results.

The additional TimesFM samples were frozen before prediction after the leading-token correction. Lower bounds exceed 0.001 for both distances and zero for directional agreement; all six Holm-adjusted p-values are 0.0012 (Table 4). The AR(1) predictions also serve the exact-risk study; Appendix B details sample separation.

Table 4: Additional TimesFM attention tests. Entries give estimates [one-sided 95% lower bounds]. The reference values are 0.001 for Δ_{mask} and Δ_{back}, and zero for ρ. The listed Holm-adjusted p-value applies to each of the three tests.

Data	Δ_{mask}	Δ_{back}	ρ	Holm p
AR1	0.0441 [0.0413]	0.0402 [0.0373]	0.9839 [0.9804]	0.0012
Traffic	0.0897 [0.0632]	0.0827 [0.0580]	0.9951 [0.9933]	0.0012

4.3 Revision direction explains the change in expected loss

Table 5 gives exact AR(1) risk changes. Moirai's natural boundary change is -0.023973, or -2.40% of $\bar{v}_H$. Subtracting the median nearby non-boundary change gives a local contrast of -0.024702, or -2.47%, as defined in Equation 7. Chronos-Bolt and TimesFM have local changes of 0.0032 and 0.0029; the equal-model average is -0.0062. Moirai's negative visibility effect shows that reading the new token reduces expected loss relative to the masked prediction.

Table 5: Exact AR(1) risk changes in units of $\bar{v}_H$. Local risk change is Δq_{local} from Equation 7; visibility effect is $q(\hat{\mathbf{y}}^{+}) - q(\hat{\mathbf{y}}^{\text{mask}})$. Negative local values indicate a risk change below the median nearby non-boundary change; negative visibility values indicate lower loss with the token visible. Brackets give two-sided 95% confidence intervals.

Model	Local risk change	Visibility effect
Chronos-Bolt	$+0.0032$ $[+0.0009, +0.0055]$	$+0.0019$ $[-0.0005, +0.0043]$
TimesFM	$+0.0029$ $[+0.0012, +0.0049]$	$+0.0027$ $[+0.0011, +0.0046]$
Moirai	-0.0247 $[-0.0317, -0.0176]$	-0.0241 $[-0.0309, -0.0174]$
Equal-model average	-0.0062 $[-0.0087, -0.0037]$	-0.0065 $[-0.0088, -0.0040]$

Let $\mathbf{e} = \hat{\mathbf{y}}^0 - \mathbf{m}$ be the pre-boundary error relative to the conditional mean in Equation 5, and let $\boldsymbol{\delta} = \hat{\mathbf{y}}^+ - \hat{\mathbf{y}}^0$ be the revision. Expanding squared errors gives

$$q(\hat{\mathbf{y}}^+) - q(\hat{\mathbf{y}}^0) = \underbrace{\frac{\|\boldsymbol{\delta}\|_2^2}{H\bar{v}_H}}_{\text{revision energy}} + \underbrace{\frac{2\langle \mathbf{e}, \boldsymbol{\delta}\rangle}{H\bar{v}_H}}_{\text{error alignment}}. \tag{8}$$

Revision energy is the nonnegative squared size of the prediction change. Error alignment measures whether that change follows or opposes the existing error. A sufficiently negative alignment term lowers expected loss. For the local contrast, we rank the four neighboring transitions by their total risk changes and select the two middle transitions. We subtract the average energy and alignment components of these same transitions from the boundary components. Their sum therefore remains exactly equal to the locally adjusted risk change. Appendix D gives the full derivation.

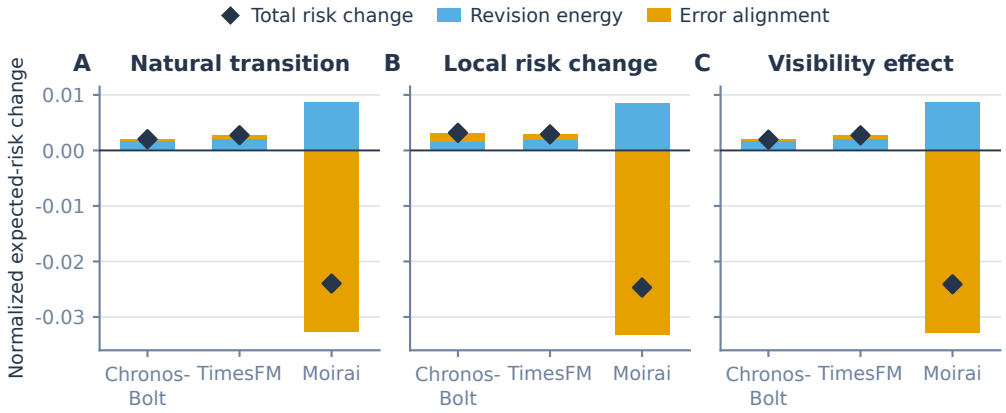

Figure 4: **Error alignment explains the expected-loss change.** Blue bars show revision energy, orange bars show error alignment, and diamonds show their sum. A: the natural $B \to B+1$ transition. B: the same comparison after subtracting the local background. C: the masked-to-visible comparison. All values use the variance scale $\bar{v}_H$. Table 15 lists the components numerically.

Figure 4 and Table 15 show that Moirai has the largest revision energy and a negative alignment term about four times as large. Its forecast therefore moves toward the conditional mean. The same pattern appears when comparing masked and visible predictions. The expected-absolute-error results in Table 6 preserve the three models' risk-change directions, connecting the exact-risk analysis to the absolute-error measure used for real data. For Gaussian futures, let $\sigma_h = \sqrt{v_h}$, $z_h = |\hat{y}_h - m_h|/\sigma_h$, and let Φ be the standard normal cumulative distribution function. Then

$$\mathbb{E}[|x_{t+h} - \hat{y}_h| \mid x_{\le t}] = \sigma_h \sqrt{\frac{2}{\pi}}\, e^{-z_h^2/2} + |\hat{y}_h - m_h|\{2\Phi(z_h) - 1\}. \tag{9}$$

We divide the horizon-averaged expected absolute error by $H^{-1}\sum_{h=1}^{H} \sigma_h \sqrt{2/\pi}$ and apply the same local and visibility contrasts. Appendix D gives further details.

Table 6: Exact expected absolute-error changes, normalized by the conditional-mean forecast's expected absolute error. Comparisons match Table 5: local change subtracts the median nearby non-boundary change; visibility compares visible and masked forecasts. Brackets give two-sided 95% confidence intervals.

Model	Local change	Visibility effect
Chronos-Bolt	$+0.00146$ $[+0.00043, +0.00250]$	$+0.00092$ $[-0.00011, +0.00191]$
TimesFM	$+0.00131$ $[+0.00056, +0.00218]$	$+0.00124$ $[+0.00050, +0.00209]$
Moirai	-0.01059 $[-0.01336, -0.00779]$	-0.01033 $[-0.01303, -0.00764]$

4.4 Real-data outcomes depend on the model and dataset

Table 7 reports local MASE changes on the separate Traffic and Electricity samples. Moirai's boundary error changes exceed the median nearby non-boundary change by 0.0391 and 0.0355, respectively. Chronos-Bolt's changes are close to zero, and TimesFM's estimates are negative on both datasets. Figure 5 places these outcomes alongside the attention measurements and exact AR(1) risks. Moirai has the largest attention effect in every dataset, while the loss direction changes between AR(1) and the real datasets. These results connect a repeatable representation effect to outcomes that depend on both the forecasting model and the data.

Table 7: Local MASE changes on separate Traffic and Electricity series. Values subtract the median nearby non-boundary loss change. Brackets are descriptive two-sided 95% confidence intervals; negative values indicate a lower error change relative to nearby non-boundary transitions.

Model	Data	Local MASE change [95% CI]
Chronos-Bolt	Traffic	$+0.0006$ $[-0.0008, +0.0025]$
Chronos-Bolt	Electricity	-0.0004 $[-0.0018, +0.0009]$
TimesFM	Traffic	-0.0042 $[-0.0084, +0.0001]$
TimesFM	Electricity	-0.0049 $[-0.0083, -0.0014]$
Moirai	Traffic	$+0.0391$ $[+0.0292, +0.0497]$
Moirai	Electricity	$+0.0355$ $[+0.0185, +0.0552]$

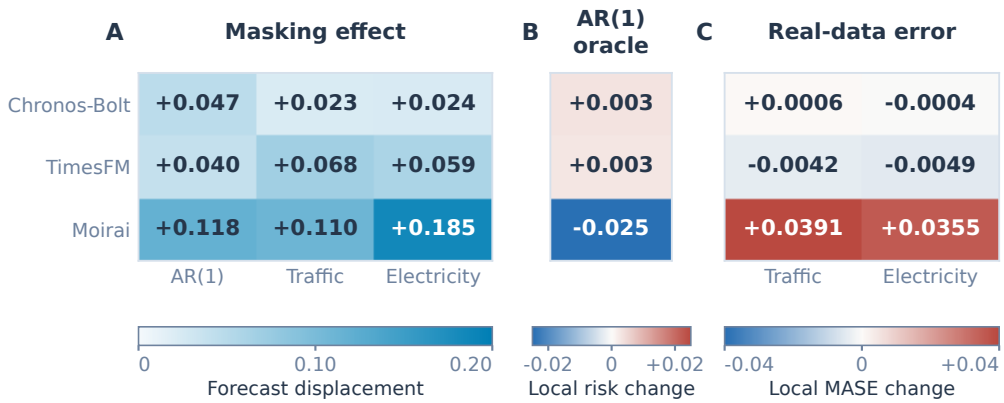

Figure 5: **The same attention mechanism has different accuracy effects.** A: the effect of masking, Δ_{mask}, in the original attention study. B: local expected-loss changes on AR(1). C: local MASE changes on the separate real-data samples. Panel A uses a sequential color scale starting at zero. In B and C, red indicates a positive local contrast and blue a negative one; each contrast is relative to nearby non-boundary transitions. Panels B and C use different loss scales; each is defined in Section 3.4. Tables 13, 5, and 7 provide the corresponding values and uncertainty.

5 Conclusion

A one-observation extension can create a new token and sharply revise a fixed-patch model's forecast. Experiments across Chronos-Bolt, TimesFM, and Moirai connect these revisions to the patch layout and to the new token's attention visibility. The exact-risk analysis explains their accuracy effects through the size and direction of the revision relative to the model's existing error. Together, these results provide a direct way to measure representation changes and their forecasting consequences. Future work can extend the analysis to multivariate forecasts, predictive distributions, and models with overlapping or adaptive patches.

AI USE STATEMENT

Generative AI tools assisted with experiment implementation, literature search, figure preparation, and writing. The authors are responsible for the experimental results, mathematical claims, citations, and final text.

ETHICS STATEMENT

This study examines publicly released forecasting models using generated time series and public benchmark data. It involves no human subjects or deployment interventions.

REPRODUCIBILITY STATEMENT

Appendix A identifies the datasets, model versions, and sample allocations. Appendices B–D give measurement details, full result tables, statistical procedures, and the expected-loss calculation.

REFERENCES

Sachith Abeywickrama, Emadeldeen Eldele, Min Wu, Xiaoli Li, and Chau Yuen. Entropy guided dynamic patch segmentation for time series transformers. *arXiv preprint arXiv:2509.26157*, 2025.

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

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'24, 2024.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan Ö. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25, 2025.

Ruijin Hua, Zichuan Liu, Kun Zhang, and Yiyuan Yang. Diversified scaling inference in time series foundation models, 2026. URL https://arxiv.org/abs/2601.17376.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22, 2006.

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

Manya Pandey, Dhruv Kumar, Murari Mandal, and Saurabh Deshpande. GITCO: Gated inference-time context optimization in TSFMs. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://openreview.net/forum?id=AKHqrrUdWD.

Debdeep Sanyal, Aaryan Nagpal, Dhruv Kumar, Murari Mandal, and Saurabh Deshpande. time2time: Causal intervention in hidden states to simulate rare events in time series foundation models, 2025. URL https://arxiv.org/abs/2509.05801.

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'24, 2024.

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

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, 2021.

Federico Zucchi, Yi Xie, Chao Zhang, Keyuan Luo, Thomas Lampert, and Ziyue Li. Adaptive patching is harder than it looks for time-series forecasting, 2026. URL https://arxiv.org/abs/2606.04074.

A Data and Model Details

The real datasets are downloaded from the Time-Series-Library collection maintained by THUML. Table 8 gives the sampling intervals and number of series. Each measured variable is treated as a separate univariate series; timestamps are excluded from model inputs. ETTh1 and ETTm1 are electricity-transformer datasets (Zhou et al., 2021); the remaining datasets use the versions distributed with the same benchmark library (Wu et al., 2023).

Table 8: Real datasets from the Time-Series-Library benchmark collection. Each measured variable is evaluated as a separate univariate series.

Dataset	Sampling interval	Series
ETTh1	1 hour	7
ETTm1	15 minutes	7
Electricity	1 hour	321
Traffic	1 hour	862
Weather	10 minutes	21

ETTh1 uses its first 8,640 observations for training, the next 2,880 for validation, and the next 2,880 for testing. ETTm1 uses 34,560, 11,520, and 11,520 observations in the same order. Electricity, Traffic, and Weather use chronological 70%/10%/20% splits. For a training segment of length T and seasonal lag s, the fixed normalization scale is

$$a = \frac{1}{T-s} \sum_{i=s+1}^{T} |x_i - x_{i-s}| + 10^{-8}. \tag{10}$$

Here i indexes training observations and s is the number of observations in one season. We use $s = 24$ for ETTh1, Electricity, and Traffic, $s = 96$ for ETTm1, and $s = 144$ for Weather.

The generated series have 8,192 observations after discarding 200 initial values. Their first 4,096 observations define the training scale. AR(1) uses coefficient 0.8 and independent Gaussian innovations with standard deviation one. SeasonalAR combines the previous value with weight 0.6 and the value 24 steps earlier with weight 0.3, plus Gaussian noise with standard deviation one. The sinusoidal generator adds a period-24 sine wave of amplitude one, a period-168 sine wave of amplitude 0.5, a linear trend of 0.002 per step, and Gaussian noise with standard deviation 0.2. The scale uses $s = 1$ for AR(1) and $s = 24$ for the other generators.

The model weights are Chronos-Bolt-small, TimesFM 2.5 200M, and Moirai 1.1 R-small. Chronos-Bolt returns its median quantile, TimesFM returns its point output, and Moirai returns the mean of its predictive distribution. The context caps in Table 1 describe the experiment settings. The frozen checkpoint revisions are listed in Table 10. The experiments use Chronos Forecasting 2.3.1, TimesFM 2.0.2, Uni2TS 2.0.0, and PyTorch 2.11.0+cu128. TimesFM uses the frozen `primary_recommended_raw_q` configuration: input normalization, the continuous quantile head, and flip invariance are enabled; positivity inference and quantile-crossing correction are disabled. The controls in Table 2 vary normalization and flip averaging. Moirai uses the `distribution_mean` point functional; the original runs obtain this mean from the analytic distribution interface. The configured sample-mean fallback budgets are 512 for the original experiments and 1,024 for the supplemental experiments.

Table 9: Sample allocation and context grids. A cutoff is a forecast time. The local study uses up to 128 series per dataset; all seven ETT series and all 21 Weather series are available.

Study	Series and cutoffs	Context lengths
Dense sweep (Figure 1C)	128 AR(1) series; one cutoff each	Every integer from 192 to 768
Patch-size study and Figure 2	32 series per generated dataset; one cutoff each	Every integer from 192 to 512
Local study	Up to 128 series per dataset; up to 32 cutoffs each	$B-2,\ldots,B+3$, $B=256,512,1024$
Original attention study	64 series per AR(1), Traffic, Electricity; eight cutoffs each	$B, B+1$, $B=256,512$
Real outcomes	Separate 64 Traffic and 64 Electricity series; eight cutoffs each	$B-2,\ldots,B+3$, $B=256,512$
Exact AR(1) risk	128 additional series; eight cutoffs each	$B-2,\ldots,B+3$, plus masked $B+1$
Additional TimesFM study	Same 128 AR(1) series plus 64 new Traffic series; eight cutoffs each	$B, B+1$, $B=256,512$

Table 10: Frozen model checkpoint revisions.

Model	Revision
Chronos-Bolt	`772f3d25d38aec6d914c8949dab4462e2d46f5d8`
TimesFM	`1d952420fba87f3c6dee4f240de0f1a0fbc790e3`
Moirai	`0c24ab99db2c1a70ea2a0fc03bf113329772ac64`

B Measurement and Attention Details

The input observation masks distinguish observed values from padding. Before the attention intervention, we count the observed values in each context patch to obtain K_{active} and f. Padding-only patches and future prediction tokens are excluded. The first patch containing an observation is the leading token. Table 11 shows the measured $256 \to 257$ transition.

Table 11: Observed context-mask counts at $B = 256$. The added token contains one observation in all three models. $f(L)$ is the number of observed values in the leading patch.

Model	P	$K_{\text{active}}(256)$	$K_{\text{active}}(257)$	$f(256)$	$f(257)$
Chronos-Bolt	16	16	17	16	1
TimesFM	32	8	9	32	1
Moirai	32	8	9	32	1

For example, the boundary at $B = 256$ compares $d_t(257)$ with the median of $d_t(255), d_t(256), d_t(258), d_t(259)$. These five adjacent revisions require forecasts at lengths 254 through 259. The same construction applies at every other boundary. The dataset-level BES is a median of per-series values as described in Section 3.1. The curves in Figures 1 and 2 instead show mean revisions. Their plotted means are used to visualize the shape of the effect; the BES table uses the separately computed local contrasts.

For phase ranks, the phase $r = L \bmod P$ is a context length's position within a patch cycle. We first average $d(L)$ at each phase within a cutoff, then take the median across cutoffs within each series and average the series values. Ranking these phase averages gives the ranks in Tables 12 and 3. The local ranks compare only the five sampled transitions with indices $B-1, B, B+1, B+2, B+3$; the dense patch-size study covers the full patch cycle. The two local settings with rank two are Moirai on the sinusoidal dataset at $H = 24$ and $H = 64$. Their BES intervals have positive lower bounds.

The attention intervention masks the leading key/value column throughout the relevant attention path. Chronos-Bolt applies it in encoder attention and in the decoder's attention to the encoder. TimesFM applies it in every self-attention layer, including both paths used to average predictions under sign reversal. Moirai applies it in the encoder attention over its packed token sequence. The projected

Table 12: Local boundary results across three models and eight datasets. A setting is one model–dataset–horizon combination. Phase rank one means that $L \bmod P = 1$ has the largest average revision among the five sampled local transitions, $L \in \{B-1, B, B+1, B+2, B+3\}$.

Horizon H	Settings	Phase rank one	BES lower bound > 0
24	24	23	24
64	24	23	24

input, positions, scaling state, and observed-token count are paired between visible and masked conditions. The leading token keeps its own query and its representation in the network. For each visible/masked pair, we check equality of the input and target, pre-attention tensor hash, position hash, scaler hash, active-token count, and leading-token index. The attention-mask hashes differ. The TimesFM audit checks the target leading column in all 40 attention calls, covering both the forward and sign-reversed paths.

After correcting TimesFM's leading-token index to locate the first active patch after left padding, we freeze the additional series-disjoint sample before prediction. These series are unused during implementation checks. The additional TimesFM experiment uses AR(1) series shared with the exact-risk study and a separate sample of Traffic series. Neither sample shares a series with the original attention or real-outcome studies. For the original Chronos-Bolt–Electricity measurements, 13 of 1,024 paired comparisons have an undefined ρ because one revision vector has zero norm. Direction summaries use comparisons with both vectors nonzero; distances remain defined. The additional TimesFM comparisons have complete paired measurements.

C Statistical Details and Additional Results

The original discovery tables use 2,000 bootstrap samples of series. Each resampled series retains its cutoff-level median, and the statistic is the median across series. The reported intervals use the 2.5th and 97.5th percentiles of these bootstrap values. The mean curves in Figures 1 and 2 use the mean across series plus or minus 1.96 times its standard error. For the patch-size plots, each series contributes one cutoff.

The attention figure and loss tables use 5,000 bootstrap samples. Series are resampled with replacement, and cutoffs are resampled within each selected series. Paired model conditions and boundary comparisons remain together. Two-sided intervals use the 2.5th and 97.5th percentiles. The additional TimesFM tests in Table 4 use the fifth percentile as a one-sided 95% lower bound. For each reference value, the reported bootstrap p-value is one plus the number of resampled averages at or below the reference, divided by 5,001. Holm correction adjusts all six tests together to account for multiple comparisons.

Table 13: Original attention-study means. Each dataset contributes 64 series and eight cutoffs per series. The three columns correspond to Equation 4; Figure 3 gives their confidence intervals.

Model	Data	Δ_{mask}	Δ_{back}	ρ
Chronos-Bolt	AR1	0.0473	0.0445	0.9799
Chronos-Bolt	Traffic	0.0229	0.0156	0.8841
Chronos-Bolt	Electricity	0.0236	0.0144	0.7974
TimesFM	AR1	0.0405	0.0367	0.9875
TimesFM	Traffic	0.0681	0.0619	0.9909
TimesFM	Electricity	0.0594	0.0539	0.9918
Moirai	AR1	0.1179	0.1153	0.9983
Moirai	Traffic	0.1102	0.1068	0.9990
Moirai	Electricity	0.1850	0.1807	0.9983

The scaling controls in Table 2 use 32 AR(1) series each. For Chronos-Bolt, fixed scaling holds the normalization parameters constant while the context grows. For TimesFM, the normalization control disables input normalization. The sign-averaging control disables the combination of predictions from the original input and a sign-reversed input. The remaining control disables both operations.

Table 14: Controls for the boundary pattern. An artificial period groups context lengths by their remainder even though Chronos-T5 has no fixed input patch. The rank compares mean revisions across those positions.

Predictor	Comparison	Result
Seasonal naive	Eight datasets; two horizons	Maximum $d(L) = 0$ in all 16 settings
Chronos-T5	AR(1); artificial periods 16/32	Position one ranks 7 / 9
Chronos-T5	Traffic; artificial periods 16/32	Position one ranks 13 / 22
Chronos-Bolt	AR(1); fixed scaling	Boundary peak remains; Table 2
TimesFM	AR(1); preprocessing controls	Boundary peak remains; Table 2

The seasonal-naive and Chronos-T5 values in Table 14 come from their own predictions on the same context-length comparisons.

D Expected-Loss Calculation

For a fixed series and cutoff, let $\mathbf{m}$ be the conditional future mean, let $\mathbf{e} = \hat{\mathbf{y}}^0 - \mathbf{m}$, and let $\boldsymbol{\delta} = \hat{\mathbf{y}}^+ - \hat{\mathbf{y}}^0$. The future horizon is H, and $\bar{v}_H$ is its average conditional variance. Substituting these definitions into Equation 6 gives

$$\begin{aligned} q(\hat{\mathbf{y}}^+) - q(\hat{\mathbf{y}}^0) &= \frac{\|\mathbf{e} + \boldsymbol{\delta}\|_2^2 - \|\mathbf{e}\|_2^2}{H\bar{v}_H} \\ &= \frac{\|\boldsymbol{\delta}\|_2^2 + 2\langle \mathbf{e}, \boldsymbol{\delta} \rangle}{H\bar{v}_H}. \end{aligned} \tag{11}$$

The identity applies to any paired forecasts with the same conditional mean and variance. For the masked-to-visible comparison, the initial error becomes $\hat{\mathbf{y}}^{\text{mask}} - \mathbf{m}$ and the revision becomes $\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}^{\text{mask}}$.

Section 4.3 explains the shared-transition subtraction that preserves the identity for the local contrast. Table 15 reports the component means.

Table 15: Components of Figure 4. Natural compares B with $B + 1$; Local subtracts the matched non-boundary background; Visibility compares masked and visible forecasts. Each row sums to its total, up to rounding.

Comparison	Model	Energy	Alignment	Total
Natural	Chronos-Bolt	+0.001822	+0.000202	+0.002024
Natural	TimesFM	+0.002148	+0.000598	+0.002746
Natural	Moirai	+0.008752	-0.032725	-0.023973
Local	Chronos-Bolt	+0.001664	+0.001508	+0.003172
Local	TimesFM	+0.001978	+0.000921	+0.002899
Local	Moirai	+0.008569	-0.033271	-0.024702
Visibility	Chronos-Bolt	+0.001741	+0.000206	+0.001946
Visibility	TimesFM	+0.002032	+0.000694	+0.002726
Visibility	Moirai	+0.008750	-0.032856	-0.024106

We also calculate exact expected absolute error from the Gaussian future distribution. Let $Y_h = x_{t+h}$ denote the future value, $\sigma_h = \sqrt{v_h}$ its conditional standard deviation, and $z_h = |\hat{y}_h - m_h|/\sigma_h$ the absolute forecast bias in standard deviation units. The Gaussian absolute-moment identity gives Equation 9 in Section 4.3. The expectation is over the conditional future distribution, and Φ is the cumulative distribution function of a standard Gaussian. We divide the horizon-averaged error by the corresponding average of $\sigma_h\sqrt{2/\pi}$, the expected absolute error of the conditional-mean forecast. Applying the same local and visibility comparisons gives Table 6. The sign of each model's change agrees with the squared-error analysis.