When There Is Nothing to Predict: Directional Bias in Time Series Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

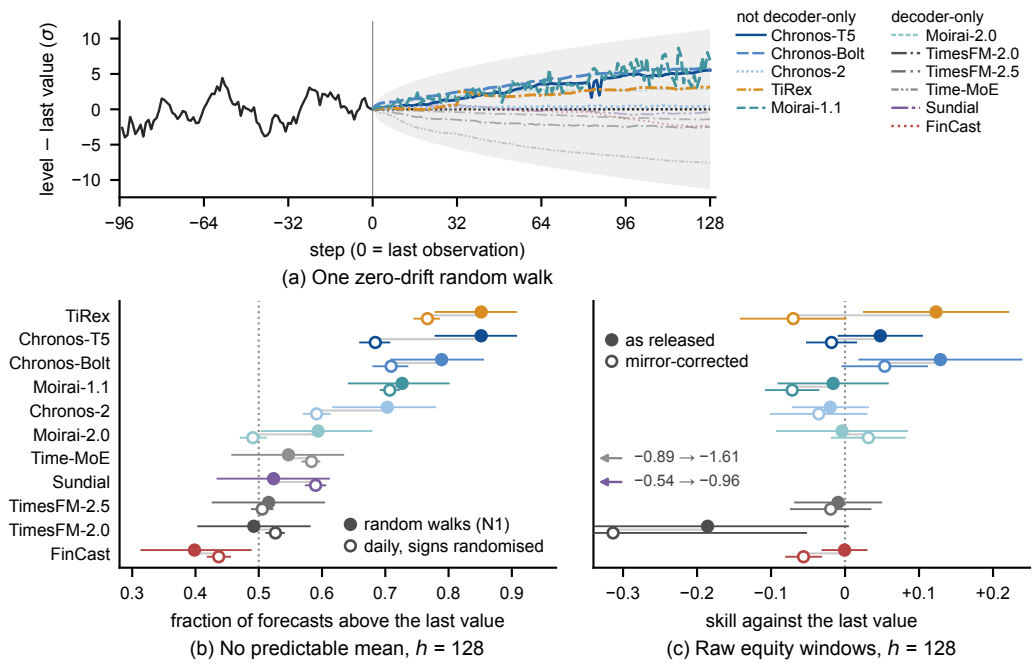
Time series foundation models forecast a new series from its context alone. We ask what they forecast when there is nothing to predict. On a martingale the last observed value is the best point forecast under squared loss, and any other forecast adds risk equal to the mean square of its departure from that value, so what a model adds can be measured against a known optimum. The part of a departure that stays the same when the history is negated cannot come from the direction of the history; removing it is the mirror correction. We find that several released models carry a directional bias that, on rising markets, looks like skill. On random walks five of the eleven models, none of them decoder-only, forecast a rise on up to 85 percent of contexts. A corpus with positive trends suffices to cause such a bias, in a decoder-only model as well. On daily prices every model loses to the last value once the return signs are randomised, and on equity windows from a rising market the mirror correction lowers 128-day skill for ten of the eleven models. On 1996–2014 windows several models time the market far better than trend rules read from the same contexts, which points to memory: public training data contain those years. The mirror correction, already built into TimesFM-2.5, lowers risk whenever past and future are symmetric under negation, and an exact identity gives its cost elsewhere; mirror augmentation during fine-tuning removes most of Chronos-Bolt's bias.

1 Introduction

Time series foundation models such as Chronos (Ansari et al., 2024), Moirai (Woo et al., 2024), TimesFM (Das et al., 2024), Timer (Liu et al., 2024b), Sundial (Liu et al., 2025b) and Time-MoE (Shi et al., 2025) are pretrained on large heterogeneous corpora and forecast a new series from its context alone. On a random walk the right forecast is known in advance: the future increment has conditional mean zero, so the last observation is the best point forecast under squared loss and any departure from it only adds risk. Prices of liquid assets are close to this case (Samuelson, 1965; Fama, 1970), which makes them a natural place to ask what these models add to a series when there is nothing to predict, and whether what they add should be read as skill.

Zero-shot evaluations on financial returns compare models where the optimum is unknown: Rahimikia et al. (2025) report negative out-of-sample R^2 for a model pretrained on daily returns, and Noguer i Alonso & Pereira Franklin (2026) find little improvement over a random walk on liquid US equities. On a rising market such comparisons cannot tell a model that reads the series from one whose bias points the right way. Yu et al. (2026) use random walks to ask which summary of the future a training loss returns, and Andreoletti (2026) split forecast error with a decomposition related to Proposition 1, taking fitted noise as the excess; we measure departures against a known optimum and ask which way they point. Li et al. (2026) audit contamination by fine-tuning a model on each dataset, and language models reproduce economic series from before their training cut-off but not after it (Lopez-Lira et al., 2025); our checks for remembered history need only forecasts and public training data (Appendix J).

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.



(a) One zero-drift random walk

(b) No predictable mean, h = 128

(c) Raw equity windows, h = 128

Figure 1: **Directional bias looks like skill on a rising market.** (a) Forecasts of the eleven models 128 steps ahead on one zero-drift random walk; the last value (dotted) is optimal, and the band is $\pm 1\sigma\sqrt{h}$. (b) Share of forecasts above the last value on 128 random walks and on 1280 daily windows with randomised return signs, with 95 percent binomial and cluster-bootstrap intervals; one half is unbiased (slightly less for a median forecast on the copies, 47 percent of whose futures rise). (c) Skill on 900 raw daily equity windows, as released and after the mirror correction (Figure 3), with ± 1 cluster-bootstrap standard errors; arrows give values beyond the axis. Models in b and c are ordered by b.

We build the diagnostic on a known optimum (Section 2; Figure 2 outlines the analysis): under a martingale null the excess squared risk of a forecast equals its mean squared departure from the last value (Proposition 1). The findings are:

1. **Five models forecast a rise when there is nothing to predict.** Every model departs from the last value and so loses to it; the five that are not decoder-only also depart upward, on up to 85 percent of random walks (each $p < 10^{-4}$ after a Holm correction), and the upward departures persist across volatility clustering, heavy tails, price levels, model sizes and context lengths (Sections 3 and 4).
2. **A corpus with positive trends suffices to set the sign.** A masked encoder and a decoder-only model trained from scratch on such a corpus both acquire an upward bias, at both sizes we trained, so the decoder-only design alone does not keep a model near balance; how the released models weight such trends is not public (Section 4).
3. **On real prices, bias and skill are confounded.** With return signs randomised, all eleven models lose to the last value; on rising equities, removing the part of the departure that ignores the direction of the history lowers 128-day skill for ten of the eleven models. In a falling-market test with predictions fixed in advance, the removal helps Chronos-T5 and Moirai-1.1 where the market fell, while four other models time the 1996–2014 market in a way that points to memory (Section 5).
4. **Removing the bias is safe only under symmetry.** The mirror correction lowers risk by the mean square of what it removes when the data are symmetric (Proposition 2), which it did in 135 of 176 synthetic cells; an assumption-free identity gives its cost elsewhere, and mirror augmentation during fine-tuning removes most of Chronos-Bolt's bias (Section 6).

Section 7 derives three checks for evaluating forecasters on financial series.

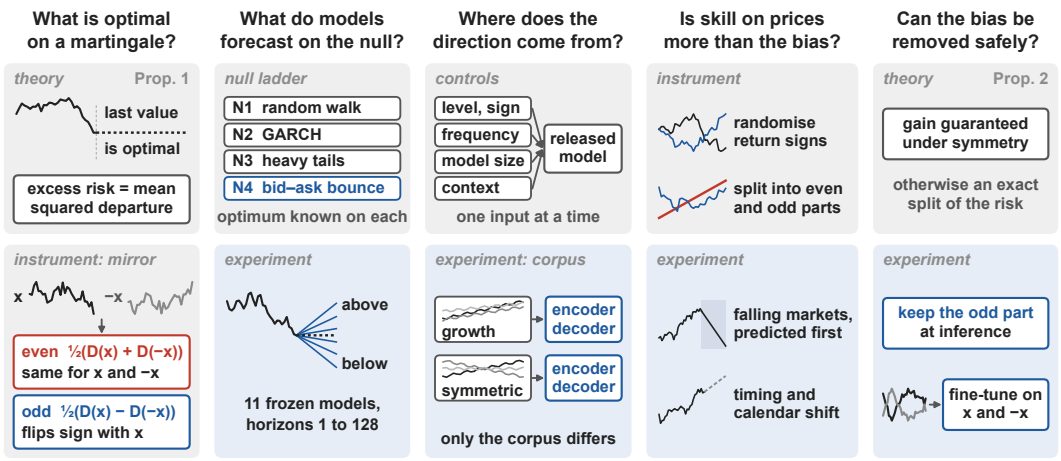


Figure 2: **How the analysis proceeds.** Columns follow Sections 2–6; grey panels hold theory and instruments, blue panels experiments; D: departure from the last value; curves are schematic.

2 Measuring Departures against a Known Optimum

2.1 Excess Risk under a Martingale Null

Let g_θ map a context $x_{1:t}$ to a forecast $\hat{y}_{t+1:t+H}$, with its pretrained weights fixed. We score its point forecast; interval width and coverage are reported separately (Table 7), since a model can scale its intervals to the volatility while its point forecast moves away from the optimum.

Definition 1 (Martingale null) *A price process $\{y_t\}$ satisfies the martingale null with respect to a filtration $\{\mathcal{F}_t\}$ if $\mathbb{E}[y_{t+h} \mid \mathcal{F}_t] = y_t$ for every horizon $h \geq 1$, with $\mathbb{E}[y_t^2] < \infty$.*

The squared-loss-optimal forecast under this null is the last value y_t, also called persistence.

Proposition 1 (Excess risk under the martingale null) *Let $\{y_t\}$ satisfy Definition 1 and let $\hat{y}$ be a square-integrable, $\mathcal{F}_t$-measurable forecast. Then*

$$\mathcal{R}(\hat{y}) := \mathbb{E}[(y_{t+h} - \hat{y})^2] = \underbrace{\mathbb{E}[(y_{t+h} - y_t)^2]}_{\textit{irreducible}} + \underbrace{\mathbb{E}[(\hat{y} - y_t)^2]}_{\textit{departure}}. \tag{1}$$

The cross term vanishes by conditioning on $\mathcal{F}_t$ (Appendix A). The identity prices a departure whatever its cause: a drift or AR(1) fitted to the context alone also departs and loses under the null (Table 1), as can a frozen model that reacts to chance patterns. We report the root mean squared departure $\sqrt{\mathbb{E}[(\hat{y} - y_t)^2]}$, which sets the excess risk, and the signed mean $\mathbb{E}[\hat{y} - y_t]$, which gives the direction and can be zero while the excess risk is large. These nulls diagnose behaviour (scope: Appendix H).

2.2 Null Processes and Controls

N1–N3 change only the innovations, and the last value stays optimal: N1 is a Gaussian random walk, N2 has GARCH$(1, 1)$ returns, whose *variance* is highly predictable while the mean is not, and N3 has Student-t innovations with $\nu \in \{3, 5\}$. N4, the bid–ask-bounce model of Roll (1984), is the positive control, on which the last value is visibly suboptimal. Its observed price is $p_t = m_t + \frac{s}{2} q_t$, where m_t is an efficient random walk, $q_t = \pm 1$ i.i.d. trade directions and s the spread, so its increments have lag-one covariance $-s^2/4$ and the price is not a martingale. We score departures against the conditional mean $p_t - \frac{s}{2}\mathbb{E}[q_t \mid p_{1:t}]$, computed by a two-state filter; Appendix B.1 distinguishes it from the optimal linear MA(1) forecast.

Controls. A departure can also come from sampling, value encoding or input scaling. A sample mean adds its Monte Carlo variance to the squared departure; we subtract its estimate there, but not

from skill or the mirror ratios. For Chronos-T5 we also vary the number of samples, the codec's resolution and the input scale and run a randomly initialised copy (Appendix C.2); spectra separate a smooth drift from seasonal peaks (Appendix D.3); and where a model takes a frequency input we change only that input (Appendix D.2).

2.3 Models, Readouts and Samples

We evaluate the eleven released models of Table 2 (unqualified, Chronos and Moirai mean Chronos-T5 and Moirai-1.1) and all five sizes of Chronos-T5, with the 46M model as the default. Each model sees one series at a time with its weights fixed; our decoding loop for Time-MoE closely reproduces its published ETTh1 results (Table 4; horizons: Appendix B.4). We score each model's own point output, the mean of $S{=}100$ sampled paths, a median or a point head (Table 2); on the symmetric additive nulls the optimum of each is the last value.

The ladder starts at level 100 with one-percent steps and has 512 series per rung, three seeds, a context of 512 and a horizon of 16. Departures are in units of one step's σ, and skill is $1-\mathcal{R}(\hat{y})/\mathcal{R}(y_t)$, zero for the last value (Appendix B.3). The direction experiment uses 128 fresh N1 series and the correction experiment 128 per rung (the correction sample), both forecast 128 steps ahead (samples: Appendix B.3). Standard errors come from a bootstrap over series on synthetic data and over 37 clusters of daily windows that share a market (exchange rates or equities) and an end date; the clusters treat markets and adjacent blocks as independent.

2.4 Mirror Decomposition

A model that forecasts a rise both after a history and after its negation carries a bias that does not come from the direction of the history (Figure 3a). Let $D(x) = g(x) - x_t$ be the departure at a fixed horizon. The *mirror* negates every increment of the context; the part of D that stays the same under it is the *even* part, and the part that changes sign with the history the *odd* part. The *sign* group negates each increment independently. All four ladder processes are mirror-invariant, N1–N3 are also sign-invariant and N4 is not; the martingale property implies neither symmetry. For a finite group $\mathcal{G}$ acting on contexts and futures, the orbit average and the corrected forecast are

$$a(x) = \frac{1}{|\mathcal{G}|}\sum_{\tau\in\mathcal{G}}\big[g(\tau x) - (\tau x)_t\big], \qquad \hat{y}^{\mathrm{c}} = g(x) - a(x). \tag{2}$$

For the mirror, a is the even part and $D - a$ the odd part; we call $\rho_{\mathrm{even}} = \mathbb{E}[a^2]/\mathbb{E}[D^2]$ the even-energy ratio. Chronos-T5 forecasts a rise after the history and, less steeply, after its mirror, so its even part is an upward drift (Figure 3b). Applied to a real window, the sign group gives a sign-randomised copy: the same absolute returns with independent random signs, a martingale on which the last value is optimal (Figure 3c). This is standard group averaging (Chen et al., 2020; Lyle et al., 2020); for the mirror it costs one extra forecast, and TimesFM-2.5 already averages over the additive mirror by default. For the sign group we average over K random draws (Algorithm 1).

Proposition 2 (Symmetry-averaged correction) *Let the context law be $\mathcal{G}$-invariant, let $\mathcal{I}_{\mathcal{G}}$ contain the information in the context orbit, and assume $\mathbb{E}[Y \mid \mathcal{I}_{\mathcal{G}}] = 0$, where $Y = y_{t+h} - y_t$. For square-integrable Y and D:*

(i) $\mathbb{E}[\hat{y}^{\mathrm{c}} - y_t \mid \mathcal{I}_{\mathcal{G}}] = 0$;
(ii) $\mathcal{R}(\hat{y}^{\mathrm{c}}) = \mathcal{R}(\hat{y}) - \mathbb{E}[a^2]$;
(iii) without either assumption, letting $m_h(x) = \mathbb{E}[Y \mid x]$,

$$\mathcal{R}(\hat{y}^{\mathrm{c}}) - \mathcal{R}(\hat{y}) = -\underbrace{\mathbb{E}[a^2]}_{\text{removed energy}} + \underbrace{2\mathbb{E}[m_h a]}_{\text{alignment}} - \underbrace{2\mathbb{E}[(D-a)a]}_{\text{odd–even cross term}}. \tag{3}$$

For $D \equiv c$, this reduces to $c(2\mathbb{E}[m_h] - c)$.

Proofs, the cost of estimating a from K draws and numerical checks: Appendix A.

The conditions hold on every rung and on the sign-randomised copies; N4's optimal departure is odd and is kept (Appendix A). Outside them part (iii) gives the exact risk change, and neither part identifies where a departure comes from: a drift fitted to the context is odd yet fits noise under the null, and a known positive drift is even yet useful (Equation 5).

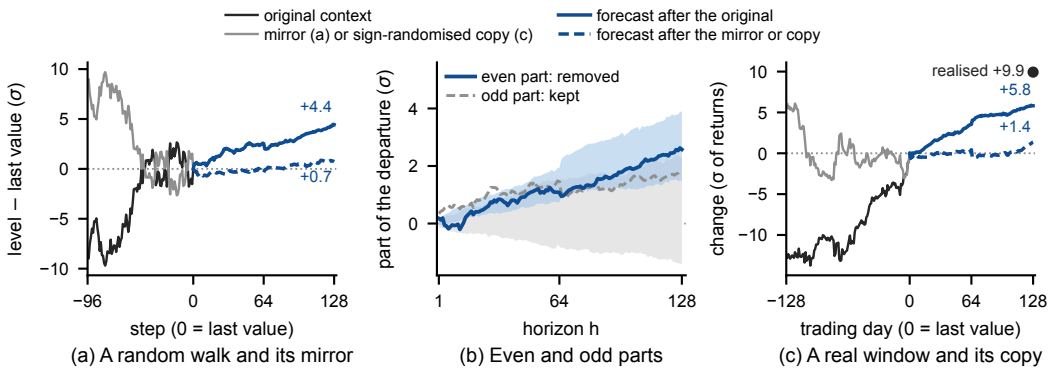


Figure 3: **A forecast that rises after a history and after its mirror carries a directional bias** (Chronos-T5; examples chosen by rule, Appendix C.1). (a) A zero-drift random walk and its mirror; the last value (dotted) is optimal. (b) The even part, the mean of the two departures, is what the mirror correction removes; the odd part, half their difference, is kept (lines: this context; bands: middle half of 128 contexts). (c) A real window (the US market, context ending 2017-11-30) and a copy with the same absolute daily returns and random signs.

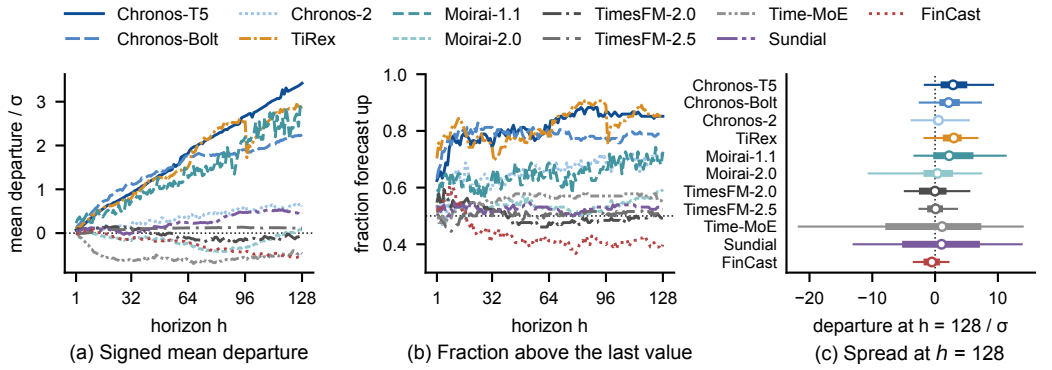


Figure 4: **On random walks the five models that are not decoder-only drift upward, more so at longer horizons.** The same 128 zero-drift random walks (N1) for all eleven models, departures in σ units; (c) median, interquartile range and 5th–95th percentiles; dotted: zero and one half.

3 What Do the Models Forecast When There Is Nothing to Predict?

Every model departs from the last value and has negative skill on every null rung (Tables 5 and 6; Figure 1a): on N1 at h=1 the departure ranges from 0.15σ for TimesFM-2.5 to 0.90σ for Moirai-1.1, and volatility clustering and heavy tails change it little. With the Monte Carlo term removed, Chronos-T5's stays between 0.21 and 0.23σ while the codec's resolution varies 750-fold (Appendix C.2). On N4, where the conditional mean is predictable and departing pays, the models capture 26 to 65 percent of its skill at h=1.

Five models lean upward. Beyond the risk that every model adds, five models add a consistent upward tilt (Figures 1b and 4; Table 8 in Appendix D.1). At h=128 on N1, the five models that are not decoder-only forecast a rise on 70 to 85 percent of contexts, Chronos-T5 and TiRex on 85 percent with mean departures of $+3.4$ and $+2.9\sigma$; the five general-corpus decoder-only models do so on 49 to 59 percent, and FinCast on 40 percent. With 128 independent series per model, exact binomial tests put these five below $p = 10^{-4}$ after a Holm correction over the eleven models, and none of the other six differs from one half after correction (Appendix D.1); we call the five the upward-biased models. For TiRex the bias is unrelated to the recent trend: its departure is almost uncorrelated with the recent slope (-0.11).

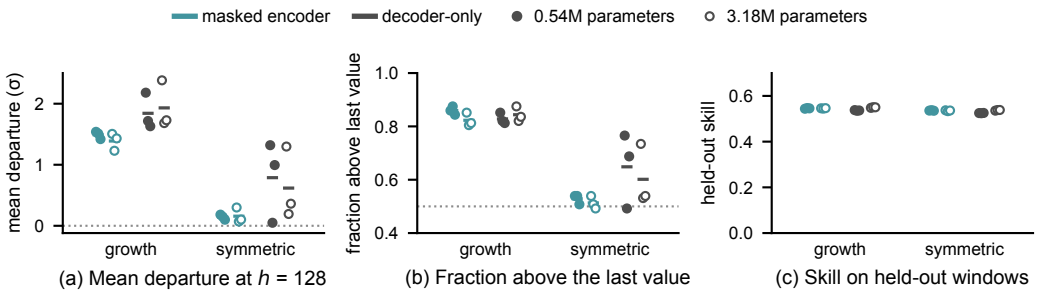


Figure 5: **A corpus with positive trends is enough to make either design forecast a rise.** A masked encoder and a decoder-only model trained from scratch on a growth or a symmetric corpus, at two sizes (dots: three seeds; dashes: their mean), scored on the 128 N1 random walks (a, b) and on held-out corpus windows (c).

4 Where Does the Direction Come From?

We test four explanations: the input's level or sign, a declared frequency, too little capacity or context, and the training corpus. Architecture, corpus, loss, normalisation and output change together across released checkpoints, so only the last test varies one factor alone. Loss and point output do not follow the split: Chronos-Bolt, Chronos-2 and TiRex lean upward and Moirai-2.0 does not, all four trained with quantile loss and read at the median (Appendix B.2).

Level, size and context do not remove the bias; a declared frequency can. The bias needs neither a rising history nor a positive price level. Chronos-T5 still forecasts a rise on 82 percent of the mirrored contexts, with a mean departure of $+6.4\sigma$ against $+3.4\sigma$ on the actual ones (Table 8), and has a positive mean departure at six of seven price levels, including zero (Figure 9). The ten models that centre and scale their inputs change by at most 0.2σ under a level shift (Table 9). A declared frequency, where a model takes one, does matter: FinCast forecasts a fall when a series is declared daily and a rise when the same values are declared weekly or yearly, and Moirai-1.1 changes with patch size (Figure 8). Chronos-T5's departures are smooth drifts, not seasonal peaks (Appendix D.3). Across its five sizes, skill on N1 shows no trend and the largest model forecasts a rise most often (Figure 10). The longest context lowers the mean departure by 30 percent for Chronos-T5 to two thirds for Moirai-1.1, but not for TiRex, and every upward-biased model still forecasts a rise more often than a fall there (Table 10).

A corpus with positive trends is enough. To vary the corpus alone, we train two small models from scratch, a masked encoder and a decoder-only model with 0.54M parameters each, on two synthetic corpora that mix seasonal, trending, random-walk and bounce series. In the growth corpus the trends and drifts have mean $+0.02$ per step; in the symmetric corpus they have mean zero. Both models normalise the context in the same way and are trained with squared error, three seeds per condition; they also differ in attention, loss positions and decoding (Appendix E.1).

Trained on the growth corpus, both models forecast a random walk to rise (86 and 83 percent of N1 contexts) while beating the last value on held-out corpus windows and on N4 (Figure 5; Table 11). On the symmetric corpus the mean departure falls from $+1.49$ to $+0.14\sigma$ for the encoder and from $+1.84$ to $+0.79\sigma$ for the decoder, whose seeds vary widely (standard deviation 0.66σ; Appendix E.2). The decoder-only model acquires the bias as readily as the encoder, so the decoder-only design does not by itself keep a model near balance, at 0.54M parameters or at six times that size (Table 12). **A corpus with positive trends is thus one sufficient cause**. Public pretraining data have such trends in part: in the datasets released with Chronos, in LOTSA and in Time-300B the typical subset does not trend, but pooled over series 52 to 57 percent of 128-step moves are rises, almost all of the excess coming from the M4 subsets (Table 13). Whether this is the cause in the released checkpoints depends on how they weight these subsets, which is not public.

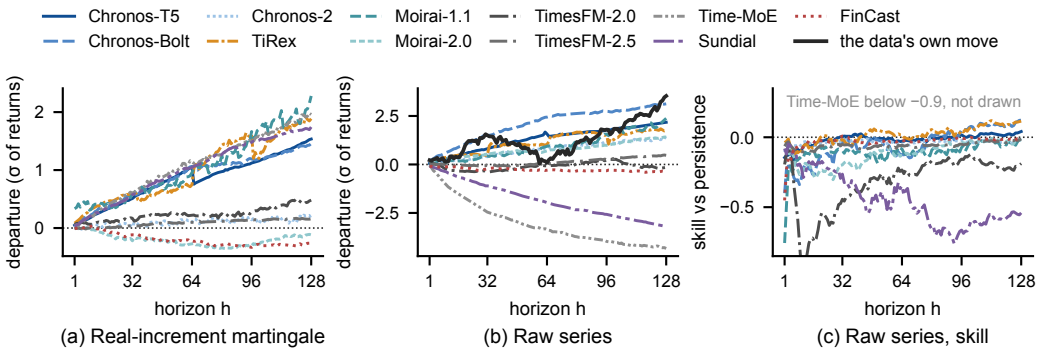


Figure 6: **On daily prices the upward departures survive randomised signs and run with the market's own rise.** 1280 windows, return units. (a) mean signed departure in σ_r units on the sign-randomised copies (four per window), where the correct value is zero; (b) the same on the raw windows, with the data's own mean move in black; (c) skill on the raw windows.

5 Is Skill on Financial Data More than the Bias?

Real prices can reward a departure, so beating the last value may mean reading the series or carrying a bias that happens to point the right way. Two tools separate the two. Randomising the signs of a window's returns keeps their magnitudes but removes any predictable mean, so a model that still departs there carries a bias. The mirror correction removes the part of a departure that stays the same when the history is negated, so skill that survives it responds to the history's direction.

With return signs randomised, every model loses. Randomising the sign of every daily return independently turns each of 1280 windows, the daily anchor, from 20 exchange rates and 50 equity portfolios (49 industries and the market) into an exact martingale with real return magnitudes (Appendix F.1). All eleven models lose to the last value at h=16 and h=128, and Chronos-T5, Chronos-Bolt, Moirai-1.1 and TiRex again forecast rises: at h=128 their mean departures are $+1.5$, $+1.5$, $+2.3$ and $+1.9\sigma_r$, with 68 to 77 percent of forecasts upward (Figures 1b and 6a), where σ_r is the context's return standard deviation; with real magnitudes the split is less clean (Table 14).

On raw prices, the skill sits in the even part. The raw windows rose, by $+3.5\sigma_r$ on average at h=128 and in 65 percent of windows, and the upward models' departures run with that rise (Figure 6b). On raw equities at h=128, removing the even part lowers skill for ten of the eleven models, for four by more than two cluster-bootstrap standard errors and for Chronos-T5 by 1.99 (Figure 1c; Table 26; nine with each window's errors scaled by its return variance, Appendix F.1). Equation 3 splits each change exactly: for Chronos-T5 its removed energy, alignment and cross term contribute $+0.040$, -0.102 and -0.004 (Table 26). The alignment term is negative for ten models, and for ten models the even part is an upward drift of 0.4 to 8.7 percent a year while the equities rose 13.6 percent a year; on exchange rates, which rose 1.8 percent a year, removing it raises skill for six models. The raw gains are real but come from the part of the departure that a rising and a falling history share: **they reflect the market's drift, not a reading of the series**.

Falling markets expose a second confound. On 1850 windows of the same portfolios ending in 1996–2014, we fixed four predictions in advance for the five upward-biased models: raw skill is lower where the market fell (P1); removing the even part lowers skill where the market rose and raises it where it fell (P2); the gap between the two regimes is larger than for the other models (P3); and every model loses on the sign-randomised copies (P4). P1, P3 and P4 hold (Table 18; P3 per model: Appendix F.5). P2 holds for Chronos-T5 and Moirai-1.1, whose even parts are upward drifts in both regimes: the correction changes their skill by -0.131 ± 0.026 and -0.110 ± 0.034 where the market rose and by $+0.064 \pm 0.050$ and $+0.208 \pm 0.062$ where it fell. It fails for Chronos-2, whose even part there points down, and for Chronos-Bolt and TiRex, because of a second confound.

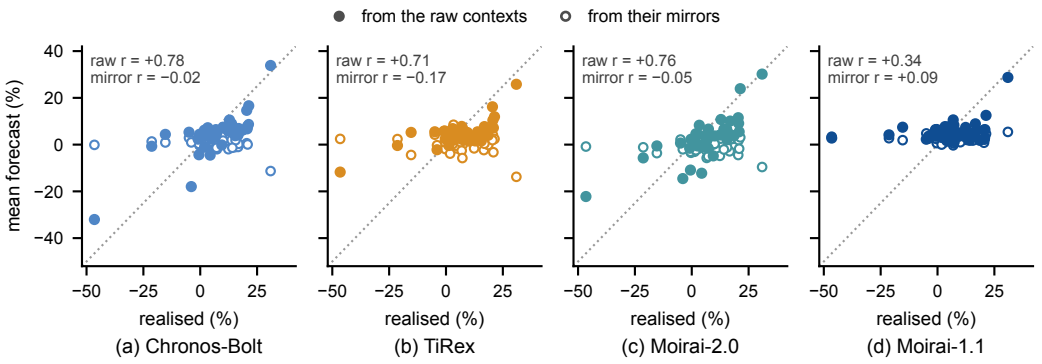


Figure 7: **Chronos-Bolt, TiRex and Moirai-2.0 time the 1996–2014 market from the actual histories only.** One point per end date (37): the mean forecast return against the mean realised 128-day return of the 50 windows ending then; dotted: forecast equals outcome. Moirai-1.1's forecasts barely move, except on the windows whose futures run from November 2008 to May 2009, where it forecasts a rise of 29 percent.

Some models time the older market. On these windows Chronos-Bolt, TiRex, Moirai-2.0 and TimesFM-2.0 time the market: across the 37 end dates their mean forecasts correlate with the realised 128-day moves at 0.71 to 0.87, against at most 0.26 for drift or momentum rules read from the same contexts (Figure 7, Table 19). A forecast that reads the direction of the context would reverse on its mirror, and one that reads its magnitudes would survive sign randomisation. For the first three models the timing neither reverses on the mirrors (-0.17 to -0.02) nor survives randomisation beyond the level of the best context rule (0.25 to 0.31 on the copies). Moving each context's end by a few weeks moves TimesFM-2.0's forecasts with the realised futures, and the other three's less clearly (Figure 11, Table 20). **Their timing points to memory of these years**, and the training data allow it: M4 daily, in several public training corpora, holds series matching the US market's daily returns on 82 percent of the 1996–2014 trading days and on none after October 2017 (Appendix F.5). Half of a remembered move lands in the even part, so there the correction cannot separate bias from memory. On the 2017–2026 windows no model clearly times the market or follows the calendar, and after 2025 the correction lowers Chronos-T5's skill from $+0.074 \pm 0.045$ to $+0.011 \pm 0.043$.

On the small-cap decile a fitted drift is as good as any model. The equal-weighted smallest size decile has real structure: positive drift and a mean lag-one return autocorrelation of $+0.32$ over the contexts. On its 120 windows from 1963 to 2026, nine models beat the last value at h=16, with skill up to $+0.391 \pm 0.070$ for Moirai-2.0, and all eleven lose on the sign-randomised copies (Table 17; skill here is normalised by each context's return variance). Fitted baselines are competitive: the context's mean return reaches $+0.306 \pm 0.059$, AR(1) $+0.327 \pm 0.060$ and AR(5) $+0.439 \pm 0.077$. Paired on the same windows, no model is ahead of the fitted drift by more than two cluster-bootstrap standard errors (the largest differences are $+0.084 \pm 0.057$ for Moirai-2.0 and $+0.079 \pm 0.108$ for Chronos-2), Time-MoE, Sundial and FinCast are behind it by more than two, and every model is behind AR(5). The largest decile and the full protocol are in Appendix F.4; no model follows the calendar on either decile (Table 20).

Structured series are forecast well. On ETTh1, with strong daily cycles, the ten general-corpus models reach skill $+0.67$ to $+0.73$ at h=16 while FinCast loses to the last value (Table 16). On one day of Nasdaq transactions no model clearly beats the last value, and neither does a fitted MA(1) (Table 15). **On financial data, beating the last value is thus weak evidence that a model reads the series.**

6 Can the Bias Be Removed Safely?

Under symmetry the correction lowers risk. On the synthetic ladder the mirror correction (Table 1) lowers risk in 135 of 176 model–rung–horizon cells and raises it in 25; TimesFM-2.5's 16 cells are unchanged, since that model already averages over the additive mirror (uncertainty in Ta-

Table 1: **The mirror correction lowers risk under symmetry and can cost skill on raw prices.** Each pair of values: raw → mirror-corrected.

Model	Daily sign-randomised			Daily raw	Synthetic	
	Departure (σ_r) $h = 128$	Skill $h = 16$	Skill $h = 128$	Skill $h = 128$	N1 skill $h = 16$	N4 capture $h = 1$
Chronos-T5	+1.54	−0.051 → **−0.020**	−0.072 → **−0.030**	**+0.041** → −0.023	−0.139 → **−0.068**	0.59 → **0.62**
Chronos-Bolt	+1.46	−0.094 → **−0.056**	−0.083 → **−0.038**	**+0.122** → +0.050	−0.098 → **−0.060**	0.62 → **0.63**
Chronos-2	+0.16	−0.037 → **−0.030**	−0.091 → **−0.067**	**−0.023** → −0.038	−0.021 → **−0.012**	0.65 → **0.67**
TiRex	+1.89	−0.052 → **−0.030**	−0.103 → **−0.059**	**+0.115** → −0.073	−0.097 → **−0.045**	0.62 → **0.67**
Moirai-1.1	+2.28	−0.207 → **−0.112**	−0.201 → **−0.080**	**−0.023** → −0.073	−0.059 → **−0.048**	0.39 → **0.41**
Moirai-2.0	−0.12	−0.082 → **−0.052**	−0.116 → **−0.074**	−0.009 → **+0.026**	**−0.037** → −0.045	0.63 → **0.64**
TimesFM-2.0	+0.48	−0.201 → **−0.152**	−0.265 → **−0.152**	**−0.188** → −0.310	−0.186 → **−0.096**	0.56 → **0.61**
TimesFM-2.5	+0.16	−0.081 → **−0.074**	−0.089 → **−0.067**	**−0.014** → −0.024	**−0.023** → **−0.023**	**0.58** → **0.58**
Time-MoE	+1.96	−1.227 → **−1.026**	−0.945 → **−0.820**	**−0.889** → −1.589	−0.554 → **−0.511**	**0.77** → 0.71
Sundial	+1.74	−0.330 → **−0.265**	−0.516 → **−0.423**	**−0.541** → −0.952	−0.179 → **−0.129**	**0.65** → **0.65**
FinCast	−0.26	−0.065 → **−0.040**	−0.053 → **−0.033**	**−0.004** → −0.056	−0.010 → **+0.009**	0.19 → **0.31**
Drift (fit)	+0.12	**−0.030** → **−0.030**	−0.187 → **−0.184**	−0.173 → **−0.133**	**−0.010** → **−0.010**	**−0.01** → **−0.01**
AR(1) (fit)	+0.12	**−0.030** → **−0.030**	−0.188 → **−0.184**	−0.174 → **−0.134**	**−0.011** → **−0.011**	**0.53** → **0.53**

The departure column is raw. Daily: 1280 windows (380 FX, 900 equities), context 512, future 128, four sign copies per window; skill uses pooled squared return errors. Synthetic: 128 scored contexts per rung; N4 capture is skill relative to the conditional mean's. Drift and AR(1) are fitted per context. Bold marks the higher value of each pair. The synthetic columns use the correction sample, whose N1 skills differ from Table 6 by sampling (Table 3). Uncertainty and full results: Appendices F.1 and G.

ble 22); on daily prices the mirror negates returns instead, and there TimesFM-2.5 keeps an even part (Appendix G.4). On N1 at h=16 the even-energy ratio runs from a tenth or less (Chronos-2, Time-MoE, Sundial) to four fifths (Moirai-1.1); for sampled outputs part of the gain is halved sampling noise (Table 25). On N4, whose optimal departure is odd and therefore kept, the mirror raises the captured share of the optimum for nine models and lowers it for Time-MoE, while the independent-sign correction lowers it for every model (Table 22). On the sign-randomised daily windows the mirror lowers risk for all eleven models at both horizons (Table 1), while sign averaging with four draws lowers skill for ten models (Tables 23 and 24).

Outside symmetry it can remove useful forecasts. On the raw windows of Section 5, pooled over both markets, Chronos-T5's raw-window skill falls from $+0.04 \pm 0.06$ to -0.02 ± 0.04, a paired change of -0.064 ± 0.033. Elsewhere its effect is small or helpful: at h=16 it raises equity skill for eight models and exchange-rate skill for seven, raises skill on the raw transaction windows for every model in Table 15, and changes the general-corpus models' ETTh1 skill by about 0.02 at most while raising FinCast's from -0.03 to $+0.05$ (Table 16). A gate that keeps the raw forecast when the context's mean return suggests the even part is useful (Algorithm 1) beats the raw forecast on equities at h=128 for only two models and has no risk guarantee (Table 27). **The mirror correction is guaranteed to lower risk only where the data are symmetric**; on drifting series, part (iii) gives its cost.

The symmetry can also be trained in. Fine-tuning the released Chronos-Bolt-small on public training windows, each replaced by its mirror with probability one half, against the same windows unchanged (three seeds each; Appendix G.8), brings its upward fraction on N1 at h=128 to 0.58 against 0.71 (0.79 as released), keeps ETTh1 skill at 0.71 against 0.70 and lowers raw-equity skill at h=128 from $+0.126$ to $+0.064$, as the predictions written before the runs said; the inference-time correction leaves $+0.060$ (Table 28).

7 Conclusion

Several released time series foundation models add a directional bias that looks like skill on a rising market. Three checks follow for forecasters on financial series: compare them with the last value, a fitted drift and, for direction, a forecast that always rises (Table 21); score them on sign-randomised copies; and report the skill that survives the mirror correction, a diagnostic whose risk guarantee holds only under symmetry. For model developers, the corpus experiment and audit suggest reporting the trend statistics of a pretraining corpus, and mirror augmentation during fine-tuning removed most of Chronos-Bolt's bias without cost on ETTh1 (Table 28). The probe needs only forecasts, so it applies to any forecaster (details and limitations: Appendix I).

ETHICS STATEMENT

The study uses synthetic series; public central-bank exchange rates, Kenneth R. French data library portfolio returns, the ETTh1 benchmark and one day of Nasdaq order-book sample data; and, for the corpus audit and the fine-tuning experiment, the public pretraining data released with Chronos, Moirai (LOTSA) and Time-MoE (Time-300B). It involves no human subjects or personal data. The reported scores assess statistical forecasts relative to persistence; they do not establish a deployable trading strategy or returns after transaction costs.

REPRODUCIBILITY STATEMENT

https://anonymous.4open.science/r/nothing-to-predict-9AD2
Synthetic processes, checkpoints and metrics are in Appendix B, data sources and windows in Appendix F, the corpus intervention in Appendix E.1, the correction in Appendix G, and proofs in Appendix A.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing.

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.

Pierre Andreoletti. Forecast collapse of transformer-based models under squared loss in financial time series. *arXiv preprint arXiv:2604.00064*, 2026.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 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.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.23719.

John M Bates and Clive W J Granger. The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468, 1969. doi: 10.1057/jors.1969.103.

John Y. Campbell and Samuel B. Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531, 2008. doi: 10.1093/rfs/hhm055.

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

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

Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417, 1970. doi: 10.2307/2325486.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. TimeGPT-1. *arXiv preprint arXiv:2310.03589*, 2023.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024. arXiv:2402.03885.

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

Hongkai Li, Shifeng Xie, Lefei Shen, Zhuo Li, Mouxiang Chen, Xiaobin Zhang, Han Fu, Jianling Sun, Xiaoxue Ren, and Chenghao Liu. TSFMAudit: Data contamination auditing in forecasting time series foundation models. *arXiv preprint arXiv:2605.26161*, 2026.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a.

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

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *International Conference on Machine Learning*, 2024b.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *International Conference on Machine Learning*, 2025b. arXiv:2502.00816.

Andrew W Lo and A Craig MacKinlay. Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies*, 1(1):41–66, 1988. doi: 10.1093/rfs/1.1.41.

Andrew W Lo and A Craig MacKinlay. An econometric analysis of nonsynchronous trading. *Journal of Econometrics*, 45(1–2):181–211, 1990. doi: 10.1016/0304-4076(90)90098-E.

Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust LLMs' economic forecasts? *arXiv preprint arXiv:2504.14765*, 2025.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020.

Richard A. Meese and Kenneth Rogoff. Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics*, 14(1–2):3–24, 1983. doi: 10.1016/0022-1996(83)90017-X.

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.

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

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

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. Lag-Llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

Richard Roll. A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance*, 39(4):1127–1139, 1984. doi: 10.1111/j.1540-6261.1984.tb03897.x.

Paul A. Samuelson. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6(2):41–49, 1965.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, 2025. arXiv:2409.16040.

Allan Timmermann. Forecast combinations. In *Handbook of Economic Forecasting*, volume 1, pp. 135–196. Elsevier, 2006. doi: 10.1016/S1574-0706(05)01004-9.

Ivo Welch and Amit Goyal. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508, 2008. doi: 10.1093/rfs/hhm014.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, 2024.

Jonathan H. Wright. Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics*, 146(2):329–341, 2008. doi: 10.1016/j.jeconom.2008.08.012.

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.

Annan Yu, Danielle C. Maddix, Boran Han, Xiyuan Zhang, Abdul Fatir Ansari, Oleksandr Shchur, Christos Faloutsos, Andrew Gordon Wilson, Michael W. Mahoney, and Bernie Wang. Understanding the implicit biases of design choices for time series foundation models. In *International Conference on Learning Representations*, 2026.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Zhuohang Zhu, Haodong Chen, Qiang Qu, and Vera Chung. FinCast: A foundation model for financial time-series forecasting. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 4539–4549, 2025. doi: 10.1145/3746252.3761261.

A Risk Identities and Proofs

A.1 Excess Risk under the Martingale Null

Write $y_{t+h} - \hat{y} = (y_{t+h} - y_t) - (\hat{y} - y_t)$ and expand:

$$\mathbb{E}\big[(y_{t+h} - \hat{y})^2\big] = \mathbb{E}\big[(y_{t+h} - y_t)^2\big] - 2\,\mathbb{E}\big[(y_{t+h} - y_t)(\hat{y} - y_t)\big] + \mathbb{E}\big[(\hat{y} - y_t)^2\big].$$

The forecast $\hat{y}$ is $\mathcal{F}_t$-measurable, so conditioning the cross term on $\mathcal{F}_t$ and applying Definition 1,

$$\mathbb{E}\big[(y_{t+h} - y_t)(\hat{y} - y_t)\big] = \mathbb{E}\Big[(\hat{y} - y_t)\,\mathbb{E}\big[y_{t+h} - y_t \mid \mathcal{F}_t\big]\Big] = 0.$$

Substituting gives (1). □

A.2 Shrinkage toward Persistence

Proposition 3 (Risk of shrinkage under the null) *Let $\{y_t\}$ satisfy Definition 1, let $\hat{y}$ be square-integrable and $\mathcal{F}_t$-measurable, and let $\pi \in [0, 1]$ be $\mathcal{F}_t$-measurable. Define $\hat{y}^\pi = y_t + \pi(\hat{y} - y_t)$. Then*

$$\mathcal{R}(\hat{y}^\pi) \;=\; \mathcal{R}(y_t) \;+\; \mathbb{E}\big[\pi^2\,(\hat{y} - y_t)^2\big], \tag{4}$$

so $\mathcal{R}(\hat{y}^\pi)$ is non-decreasing in π pointwise, is minimised at $\pi \equiv 0$, and satisfies $\mathcal{R}(\hat{y}^\pi) \le \mathcal{R}(\hat{y})$ whenever $\pi \le 1$.

Proof. Apply Proposition 1 to $\hat{y}^\pi$. Since $\hat{y}^\pi - y_t = \pi(\hat{y} - y_t)$ and π is $\mathcal{F}_t$-measurable, the departure term is $\mathbb{E}[\pi^2(\hat{y} - y_t)^2]$, which gives (4). It is non-negative, and non-decreasing in π pointwise, so it is minimised at $\pi \equiv 0$ where it vanishes and $\mathcal{R}(\hat{y}^0) = \mathcal{R}(y_t)$; and at $\pi \equiv 1$ it equals $\mathcal{R}(\hat{y})$, so any $\pi \leq 1$ gives risk no larger than the ungated forecast. □

Remark 1 (Dependence on the null) *The result assumes that persistence is optimal; outside this null, shrinking towards persistence can discard useful predictions.*

A.3 Orbit Averaging and the General Risk Identity

Two further parts of Proposition 2 cover the sign-averaged correction and the optimal forecaster.

Proposition 4 (Finite draws and orbit constancy) *Under the assumptions of Proposition 2:*

(a) estimating a from K independent uniform group draws adds $\mathbb{E}[V(x)]/K$ to the risk in Proposition 2(ii), where $V(x) = \mathrm{Var}_\tau[D(\tau x)]$. Risk is no greater than $\mathcal{R}(\hat{y})$ iff $\mathbb{E}[a^2] \geq \mathbb{E}[D^2]/(K+1)$, with strict improvement for a strict inequality;

(b) a is constant along each orbit, even without distributional assumptions. Under the stated assumptions, it vanishes for the optimal forecaster $g^\star(x) = \mathbb{E}[y_{t+h} \mid x]$.

Write $Af(x) = |\mathcal{G}|^{-1} \sum_{\tau \in \mathcal{G}} f(\tau x)$ for the orbit average of a function of the context, so that $a = AD$ with $D(x) = g(x) - x_t$. A is linear, idempotent, its range is the functions constant along orbits, and when the law of x is $\mathcal{G}$-invariant it is the conditional expectation given $\mathcal{I}_\mathcal{G}$: for $\mathcal{I}_\mathcal{G}$-measurable ϕ, $\mathbb{E}[\phi\, f] = \mathbb{E}[\phi\, f \circ \tau]$ for every τ by invariance, and averaging over τ gives $\mathbb{E}[\phi f] = \mathbb{E}[\phi\, Af]$. That AD depends on x only through its orbit is the first claim of Proposition 4(b) and needs no assumption.

(i) $\mathbb{E}[D - AD \mid \mathcal{I}_\mathcal{G}] = AD - AD = 0$.

(ii) With $Y = y_{t+h} - y_t$, expand the risk difference:

$$\begin{aligned}\mathbb{E}[(Y - D + AD)^2] - \mathbb{E}[(Y - D)^2] &= \mathbb{E}[(AD)^2] + 2\mathbb{E}[AD(Y - D)] \\ &= -\mathbb{E}[(AD)^2] + 2\mathbb{E}[ADY] - 2\mathbb{E}[AD(D - AD)].\end{aligned}$$

The last term vanishes because AD is $\mathcal{I}_\mathcal{G}$-measurable and $\mathbb{E}[D - AD \mid \mathcal{I}_\mathcal{G}] = 0$; the middle term vanishes because $\mathbb{E}[y_{t+h} - y_t \mid \mathcal{I}_\mathcal{G}] = 0$ by assumption.

For the additive mirror, the orbit-mean assumption follows from joint invariance: with $m(x) = \mathbb{E}[y_{t+h} - y_t \mid x]$, invariance of the joint law under $(x,\, y_{t+h} - y_t) \mapsto (\bar{x},\, -(y_{t+h} - y_t))$ gives $m(\bar{x}) = -m(x)$, and given the orbit $\{x, \bar{x}\}$ the two elements are equally likely, so the conditional mean of the future increment is $\frac{1}{2}(m(x) + m(\bar{x})) = 0$. For the sign group, the martingale assumption gives the zero orbit-conditional mean by the tower property; invariance of the context law under that group remains a separate requirement.

Proposition 4(a). Let $\hat{D}_K$ be the average of $D(\tau_k x)$ over K independent uniform draws $\tau_k \in \mathcal{G}$, so $\hat{y} - \hat{D}_K = \hat{y}^{\mathrm{c}} - (\hat{D}_K - AD)$. Given x the draws are independent of y_{t+h} and $\mathbb{E}[\hat{D}_K - AD \mid x] = 0$, so the cross term vanishes and the risk gains $\mathbb{E}[(\hat{D}_K - AD)^2] = \mathbb{E}[V(x)]/K$ with $V(x)$ the variance of $D(\tau x)$ over uniform τ; this uses only our own randomisation. Under invariance $V(x) = \mathrm{Var}[D \mid \mathcal{I}_\mathcal{G}]$, so $\mathbb{E}[V] = \mathbb{E}[D^2] - \mathbb{E}[(AD)^2]$, and with (ii) the risk of the estimate minus that of $\hat{y}$ is $-\mathbb{E}[(AD)^2] + (\mathbb{E}[D^2] - \mathbb{E}[(AD)^2])/K$, non-positive iff $\mathbb{E}[(AD)^2] \geq \mathbb{E}[D^2]/(K+1)$.

Proposition 4(b). For $g^\star$, $D = m$ and $Am = \mathbb{E}[m \mid \mathcal{I}_\mathcal{G}] = \mathbb{E}[y_{t+h} - y_t \mid \mathcal{I}_\mathcal{G}] = 0$.

(iii) follows from the expansion in (ii), using $\mathbb{E}[AD\,Y] = \mathbb{E}[AD\, m_h]$. The last two terms are a forecast–conditional-mean product and an odd–even cross moment. Without invariance the cross moment need not be a covariance. For $D \equiv c$, $AD = c$ and $D - AD = 0$, giving $c(2\mathbb{E}[m_h] - c)$. □

The constant benchmark and its remainder. Let $a = AD$, $c = \mathbb{E}[a]$ and $\bar{\mu} = \mathbb{E}[m_h]$. Separating second moments and covariance in part (iii) gives the exact identity

$$\begin{aligned}\mathcal{R}(\hat{y}^{\mathrm{c}}) - \mathcal{R}(\hat{y}) = c(2\bar{\mu} - c) - \mathrm{Var}(a) + 2\,\mathrm{Cov}(m_h, a) \\ - 2\mathbb{E}[(D - a)a].\end{aligned} \tag{5}$$

For a constant departure $D \equiv c > 0$ removal improves risk when $\bar{\mu} < c/2$ (for $c < 0$, when $\bar{\mu} > c/2$); for a variable departure the variance, covariance and cross terms also enter. Parity does not identify the source of a departure: under a known constant positive drift the optimal departure is even, and a drift estimated from a null context is odd but fits noise.

Prediction randomness and the two Monte Carlo quantities. The proposition treats g as a fixed function; it also holds conditionally on a fixed prediction random state shared by every transformed evaluation. The variance V/K in Proposition 4(a) comes from drawing group elements and is distinct from the sampling variance of a model's S-sample mean, whose share of the mirror's gain Appendix G.5 measures.

Remark 2 (Additive and multiplicative symmetries) *On the additive ladder, joint mirror invariance gives $m_h(\bar{x}) = -m_h(x)$, so any conditional-mean component is odd. N4 has this symmetry but not independent-sign invariance: randomising its signs destroys return dependence. Real data can violate either assumption through drift, skew or sign–volatility dependence. On such data the algebraic decomposition and part (iii) still hold, but the risk-reduction guarantee need not.*

On price levels the experiment uses simple returns and constructs $\bar{x}_i = \bar{x}_{i-1}(1 - r_i)$ or $x^{\varepsilon}_i = x^{\varepsilon}_{i-1}(1 + \varepsilon_i r_i)$. The departure is $g(x)/x_t - 1$, with squared risk in return units. Negating each return does not negate the compounded future return: the sum of the two compounded returns contains $2(e_2 + e_4 + \cdots)$, where e_j denotes the elementary symmetric polynomial of degree j in the future returns. The martingale definition does not itself impose sign symmetry. In our surrogates, magnitudes are held fixed and signs are drawn independently, making all nonempty products of distinct future signed returns have conditional mean zero. This gives both the required invariance and zero future mean. Mirror-invariant dependent returns need not satisfy the latter after compounding, so raw-window correction is assessed with part (iii).

A.4 Numerical Checks

Each identity was checked on simulated instances: Proposition 1 and its vanishing cross term, the shrinkage identity of Proposition 3 with its monotonicity and endpoints, and the linear growth of irreducible risk with h; the N4 increment autocovariance $-s^2/4$, the MA(1) forecast and the two-state filter (calibration, residuals, oddness under the mirror); Propositions 2 and 4, including the finite-K risk and the constant-departure sign change at twice the drift; and the gate statistics of Appendix G.7 under the null and a periodic alternative. All checks pass.

B Evaluation Protocol

B.1 Synthetic Processes and Reference Forecasts

The ladder separates avoidance of excess risk on N1–N3 from use of a known conditional mean on N4. Each series starts at $y_0 = 100$ with one-step scale $\sigma = 0.01|y_0|$ and is generated for the context plus forecast horizon in one draw.

N1: Gaussian random walk. The process is $y_t = y_{t-1} + \sigma\varepsilon_t$, with independent standard Gaussian innovations. Persistence is its conditional-mean forecast.

N2: GARCH volatility. N2 replaces σ by conditional volatility satisfying $\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta\sigma_{t-1}^2$, with $\alpha = 0.09$, $\beta = 0.90$ and $\omega = \sigma^2(1 - \alpha - \beta)$. Here r_t is the additive increment. The recursion starts at σ^2, its unconditional variance matches N1, and the volatility path is retained for the interval diagnostic (Remark 3).

N3: heavy-tailed innovations. Independent Student-t innovations with $\nu = 3$ or 5 are standardised to unit variance. The tails change while persistence remains the conditional-mean optimum.

N4: bid–ask bounce and two reference forecasts. The observed price is $p_t = m_t + \frac{s}{2}q_t$, where m is a Gaussian random walk with step $\sigma_m = \sigma$, $q_t = \pm 1$ i.i.d. and spread $s = 4\sigma_m$. The observed increments have lag-one covariance $-s^2/4$ (Roll, 1984). The optimal linear forecast is

$p_T + \vartheta e_T$, where ϑ is the invertible MA(1) coefficient and e_T its innovation recovered from observed increments. The coefficient satisfies $|\vartheta| < 1$ and $\vartheta/(1+\vartheta^2) = (-s^2/4)/(\sigma_m^2 + s^2/2)$. The squared-loss optimum is instead $\mathbb{E}[p_{T+h} \mid p_{1:T}] = p_T - \frac{s}{2}\mathbb{E}[q_T \mid p_{1:T}]$ for every $h \geq 1$: future directions have mean zero and m is a martingale. A two-state forward filter computes the posterior of q_T; the increment conditional on (q_{t-1}, q_t) is Gaussian with mean $\frac{s}{2}(q_t - q_{t-1})$.

On the ladder draws, the conditional mean has skill $+0.384$ versus $+0.258$ for the linear forecast. The former is nonlinear in the observed history because the hidden direction is two-valued. All N4 oracle-skill shares use the conditional mean; the linear comparison is identified explicitly when reported. N4 thus supplies a positive control: a forecaster that always returns persistence avoids null excess risk but cannot use this predictable component.

B.2 Checkpoints and Point-Output Conventions

The evaluation uses published weights and the univariate interface of eleven models. Chronos-T5 spans five sizes, 8M–710M, with small (46M) as the default. The other checkpoints and native outputs are listed in Table 2. Unqualified "Chronos" and "Moirai" refer to Chronos-T5 and Moirai-1.1, respectively. Moirai-1.1 uses its automatic patch setting unless stated otherwise. Covariates and multivariate inputs are not used, including for models that support them.

Table 2 groups models by the point forecast used. On N1–N3 the conditional mean and median coincide at y_t; on N4 and on multistep multiplicative copies, median outputs are scored against a conditional mean.

Table 2: Evaluated checkpoints and point forecasts.

Model	Params. (M)	Native forecast	Point forecast used
Sample means			
Chronos-T5	46	Token-based samples	Mean of 100 samples
Moirai-1.1	14	Mixture-distribution samples	Mean of 100 samples
Sundial	128	Flow-based samples	Mean of 100 samples
Medians			
Chronos-Bolt	47	Nine quantiles	Median
Chronos-2	119	Nine quantiles	Median
TiRex	35	Nine quantiles	Median
Moirai-2.0	11	Nine quantiles	Median
TimesFM-2.5	200	Continuous quantile head	Median
Other point outputs			
TimesFM-2.0	500	Point and nine quantiles	Mean channel (index 0)
Time-MoE	200[a]	Autoregressive point output	Greedy point forecast
FinCast	1000	Point and nine quantiles	Channel labelled "mean"

M: million parameters; [a]activated parameters. Checkpoints are small for Chronos-T5 and Chronos-Bolt, 1.1-R-small and 2.0-R-small for Moirai, and base for Sundial. TiRex's output labelled "mean" and the point outputs of Chronos-Bolt, Chronos-2 and TimesFM-2.5 equal their medians. TimesFM-2.5 uses flip averaging by default; for TimesFM-2.0 the mean channel is used.

Training targets. Chronos-T5 uses cross-entropy over 4096 bins, Moirai-1.1 a mixture negative log-likelihood, and Sundial the TimeFlow objective. Chronos-Bolt, Chronos-2, TiRex and Moirai-2.0 use quantile loss. TimesFM-2.0 uses a mean-squared-error point loss alongside quantile heads. Time-MoE uses Huber loss; FinCast uses point and quantile loss in place of squared error (Zhu et al., 2025). Time-MoE decoding is checked in Table 4.

B.3 Samples and Risk Metrics

The initial ladder uses 512 series per rung, three seeds, context 512 and horizon 16. Artefact sweeps use 256 series; direction, spectral and scale tests use 128, with their horizons and seeds stated in the respective sections. Root mean squared departure is measured in one-step σ units, and skill is $1 - \mathcal{R}(\hat{y})/\mathcal{R}(y_t)$. For sampled means, estimated sample variance divided by S is subtracted from squared departure before taking the root (Appendix C.2), and from squared error only in Table 10.

We call a fixed set of real windows an anchor: the daily anchor is the 1280 windows of Section 5, and Appendices F.2 and F.5 add a transaction anchor and a falling-market anchor. In the tables, bold marks the entries named by the rule in each table note, usually the best displayed point estimate within a stated comparison, ties included; such bold is not a significance test.

Samples behind the N1 skills. Table 3 lists the synthetic samples. Under N1, skill is $2\mathbb{E}[YD]/\mathbb{E}[Y^2] - \mathbb{E}[D^2]/\mathbb{E}[Y^2]$, with Y the realised increment and D the departure; the first term has zero mean but, at 128 series, moves with the realised paths. For Chronos-T5 at h=16 the squared-departure term is -0.052 on the direction contexts and -0.053 on the correction sample, while the first term is $+0.045$ on the direction contexts, whose paths rise by 0.93σ on average, and -0.085 on the correction sample, whose paths fall by 0.61σ. The gap between the two skills, 2.4 standard errors, comes almost entirely from the paths.

Table 3: Synthetic samples behind the reported N1 skills.

					N1 skill, h = 16	
Sample	**Series × draws**	**Context**	**Requested H**	**Used in**	**Chronos-T5**	**Chronos-Bolt**
Ladder	512 × 3	512	16	Tables 5, 6	–0.033	–0.116
Direction contexts	128 × 1	128–512	128	Section 3; Table 10[a]	+0.003[b]	–
Long-context paths	128 × 1	128–2048	128	Table 10[a]	–	–0.141
Correction sample	128 × 1	512	128	Tables 1, 22	–0.139	–0.098
Fine-tuning check	512 × 1	512	16	Table 28	–	–0.171

Context ranges are nested suffixes with fixed futures; the correction sample also forecasts the 128 mirrors, unscored. [a]Chronos-T5 and FinCast use the direction contexts, the other models the long-context paths. [b]With the Monte Carlo variance of the sample mean removed from the squared error (-0.006 without). At 128 series the skill standard errors are 0.044 for Chronos-T5 and 0.059 for Chronos-Bolt (Table 22).

B.4 Horizon Handling and Decoding Validation

Moirai-1.1 predicts each requested horizon in one pass and is instantiated separately for each, and TimesFM-2.0 is run with its three-level frequency indicator. Time-MoE (200M) is decoded greedily: each call selects the largest output head that fits the remaining horizon (8 steps for H=16, 64 for H=128) and re-encodes the full sequence, so forecasts at different requested horizons can involve different computations. The six later models give horizon-prefix-equivalent outputs except Chronos-2, whose forecasts for H=16 and H=128 on the same context differ by 0.10σ (rms); each experiment uses its own requested horizon.

Time-MoE decoding check. Our decoding loop reproduces the published zero-shot ETTh1 results (Table 4): with each channel standardised by training-set statistics, test windows at stride one, contexts 512 to 3072 and horizons 96 to 720, both this loop and Shi et al. (2025) give mean MSE 0.394 and MAE 0.419 over the four horizons. The per-window normalisation of our null experiments gives 0.422 and 0.438 on the same check.

Table 4: Time-MoE decoding under benchmark and per-window normalisation.

	h = 96 T = 512		h = 192 T = 1024		h = 336 T = 2048		h = 720 T = 3072		Average	
Evaluation	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
Published (Time-MoE large)	0.350	0.382	0.388	0.412	0.411	0.430	0.427	0.455	0.394	0.419
This loop, benchmark scaling	0.350	0.382	0.390	0.412	0.412	0.431	0.423	0.452	0.394	0.419
This loop, per-window scaling	0.354	0.381	0.410	0.426	0.480	0.480	0.444	0.464	0.422	0.438

Time-MoE large (200M activated); published values from Shi et al. (2025). Benchmark scaling uses training-set channel statistics and all test windows at stride one. T is context length; averages give equal weight to the four horizons. MSE and MAE are lower-is-better. The final row changes normalisation and is not ranked against the benchmark rows.

C Point Risk, Artefact Controls and Intervals

C.1 Departures across the Synthetic Ladder

Figure 1a shows the eleven forecasts on one random walk, context 115 of the 128 direction contexts, chosen once for its flat recent slope and a last value near its starting level. All eleven models depart from persistence and have negative skill on each null rung (Tables 5 and 6). Adding volatility clustering or heavy tails changes magnitudes but does not remove the departures: Chronos-T5's one-step departure is 0.25, 0.25, 0.24 and 0.24σ on the four null rungs.

The examples of Figure 3 are chosen by rule. Panels (a) and (b) use the context of the correction sample (Appendix G.1) whose even part and absolute odd part at h=128 are closest to their medians over the 128 contexts: Chronos-T5 forecasts a rise of 4.4σ after the history and of 0.7σ after its mirror. Panel (c) uses the market-portfolio window of the daily anchor whose first copy's departure at h=128 is closest to the median over all 3600 equity copies: the forecasts rise 5.8 and 1.4 return standard deviations, and the market rose 9.9.

Table 5: Root mean squared departure across the synthetic ladder.

	N1		N2		N3 (ν = 3)		N3 (ν = 5)		N4	
Model	h = 1	h = 16	h = 1	h = 16	h = 1	h = 16	h = 1	h = 16	h = 1	h = 16
Chronos-T5	0.25	0.72	0.25	**0.73**	0.24	**0.72**	0.24	0.74	1.67	2.36
Chronos-Bolt	0.41	1.28	0.43	1.26	0.41	1.12	0.42	1.26	1.43	2.35
Chronos-2	0.24	**0.60**	0.28	0.83	0.25	0.79	0.23	**0.66**	1.50	1.94
TiRex	0.17	0.96	0.20	1.00	**0.14**	0.85	**0.16**	0.91	1.48	2.13
Moirai-1.1	0.90	1.32	1.00	1.28	0.87	1.29	0.93	1.24	1.59	2.26
Moirai-2.0	0.32	1.27	0.34	1.08	0.26	1.23	0.30	1.32	1.51	2.05
TimesFM-2.0	0.33	1.09	0.34	1.26	0.32	1.31	0.35	1.47	1.54	2.09
TimesFM-2.5	**0.15**	0.73	**0.18**	0.94	0.15	0.90	**0.16**	0.80	1.56	2.01
Time-MoE	0.31	3.49	0.33	3.76	0.34	3.91	0.33	3.82	1.74	3.78
Sundial	0.30	1.95	0.30	2.22	0.24	1.74	0.28	1.85	1.64	3.10
FinCast	0.56	0.92	0.54	0.95	0.54	0.86	0.55	0.90	0.88	1.38

Departure is in one-step σ units; estimated Monte Carlo variance is removed for sampled means only. N1: Gaussian random walk; N2: GARCH; N3: Student-t innovations; N4: dependent microstructure control. Each rung uses 512 series and three seeds, with context 512 and requested horizon 16. Readouts follow Table 2. Bold marks the smallest departure in each N1–N3 column; N4 departures are not ranked.

Table 6: Skill against persistence across the synthetic ladder.

	N1		N2		N3 (ν = 3)		N3 (ν = 5)		N4	
Model	h = 1	h = 16	h = 1	h = 16	h = 1	h = 16	h = 1	h = 16	h = 1	h = 16
Chronos-T5	–0.084	–0.033	–0.097	–0.041	–0.041	–0.065	–0.097	–0.043	+0.204	–0.046
Chronos-Bolt	–0.160	–0.116	–0.180	–0.157	–0.204	–0.086	–0.191	–0.118	+0.221	–0.041
Chronos-2	–0.043	**–0.003**	–0.028	**–0.037**	–0.028	**–0.044**	–0.048	**–0.021**	+0.238	+0.045
TiRex	–0.034	–0.051	**–0.027**	–0.054	**–0.023**	–0.062	**–0.012**	–0.053	**+0.247**	+0.022
Moirai-1.1	–0.860	–0.137	–1.072	–0.130	–0.911	–0.141	–1.006	–0.104	+0.140	–0.080
Moirai-2.0	–0.098	–0.092	–0.103	–0.094	–0.074	–0.105	–0.081	–0.101	+0.236	**+0.056**
TimesFM-2.0	–0.070	–0.066	–0.086	–0.090	–0.072	–0.088	–0.130	–0.143	+0.237	+0.024
TimesFM-2.5	**–0.014**	–0.024	–0.028	–0.039	–0.024	–0.058	–0.023	–0.026	+0.233	**+0.056**
Time-MoE	–0.095	–0.824	–0.093	–0.900	–0.124	–1.110	–0.131	–1.001	+0.223	–0.346
Sundial	–0.070	–0.262	–0.095	–0.316	–0.050	–0.235	–0.067	–0.223	+0.240	–0.134
FinCast	–0.289	–0.065	–0.267	–0.051	–0.297	–0.046	–0.315	–0.032	+0.098	+0.009

Same samples and readouts as Table 5. Skill is $1 - \mathcal{R}(\hat{y})/\mathcal{R}(y_t)$; higher is better, with zero for persistence. On N1–N3, zero is the population optimum. On N4 at $h = 1$, conditional-mean skill is +0.384 and optimal linear skill is +0.258. Bold marks the highest skill in each column.

Departures grow with the horizon: Chronos-T5's from a quarter of a step at h=1 to three quarters at h=16 (skill -3 to -10 percent), Time-MoE's from 0.31–0.34σ to 3.5–3.9σ (skill -82 to -111 percent). FinCast's one-step departure is twice Chronos-T5's (skill -27 to -32 percent), Chronos-Bolt, which has no value-token grid, departs 0.41σ, and Moirai-1.1's one-step departure, the largest (0.90σ), depends on its automatic patch choice (Appendix D.2).

N4 reverses the role of departure: Chronos-T5 departs 1.67σ and achieves skill $+0.204$ at h=1, capturing 53 percent of the conditional mean's gain. TimesFM-2.0 captures 62, Time-MoE 58, Moirai-1.1 36 and FinCast 26 percent; the remaining six models capture 58–65 percent. Against the optimal linear forecast, Chronos-T5's share would be 79 percent.

The cross term in the sample risk expansion is at most 2.5 percent of irreducible risk on N1–N3 for ten models and 5.9 percent for Time-MoE. On N4 at h=1 it is 9–28 percent.

C.2 Sampling and Token-Resolution Controls

Sampling variance and resolution references. For a mean of S sampled trajectories the sampling variance adds to the expected squared departure; we subtract its estimate (sample variance over S) from the departure diagnostic, and a sweep over S fits $D + V/S$. Skill and mirror ratios keep the finite-S noise. Chronos's persistence round trip measures its token-grid resolution, 0.73σ at the one-percent step scale; because a mean of grid values can interpolate, the round-trip error is reported but never subtracted, and a sweep over the step scale tests whether the departure follows the grid. The level sweep (Appendix D.4) covers all eleven models; the other controls, including a randomly initialised Chronos reference, were run on Chronos-T5.

Residual departure at finer resolution. Chronos's round-trip error is 0.213σ, compared with grid spacing 0.73σ. The Monte Carlo sweep fits $D+V/S$ with $R^2 = 0.996$, giving root departure 0.242σ at $S \to \infty$, compared with 0.256σ at $S = 100$. As the step rises from 10^{-4} to 10^{-1} of the level, grid resolution changes 750-fold while the MC-adjusted departure at $h = 1$ settles at 0.21–0.23σ. Its ratio to grid spacing rises from 0.7 to 6.2 at $h = 1$ and to 47 at $h = 16$. Varying the price level over four orders of magnitude while keeping relative step scale fixed leaves that departure equal to four decimals.

The top one percent of series account for 7.4 percent of squared departure, and the randomly initialised reference departs by 191σ. **Chronos's departure is therefore not a rounding artefact of its grid**, although at daily return scales the grid still limits the shortest horizons (Appendix F.1).

C.3 Volatility Response and Interval Coverage

Remark 3 (Point and interval diagnostics) *N2 has an unpredictable conditional mean and predictable conditional variance. It tests whether a model can respond to volatility while keeping its point forecast near persistence. Models without intervals are evaluated only on their point forecasts.*

On N2, Chronos-T5's 80 percent interval width correlates with the true conditional volatility at $+0.79$ and $+0.84$ at h=1 across two seeds, and with the last in-context volatility at $+0.78$ and $+0.82$. Half of the variation in width across series is not explained by horizon. Coverage in this diagnostic is 71 percent against a nominal 80 percent. **The widths respond to volatility, but the intervals undercover**, so location, width and coverage have to be evaluated separately.

Table 7 reports coverage and width for all ten models with intervals, using S=100 sample paths for sampled outputs and the 0.1 and 0.9 heads for quantile outputs. On N1, nine models have coverage of 0.73–0.87 at h=1 and 0.74–0.84 at h=16. Their mean widths are 0.95–1.35 and 0.94–1.12 times the Gaussian reference width, respectively. Sundial undercovers more strongly: 0.25 and 0.43 coverage with relative widths 0.27 and 0.50.

Table 7: Coverage and width of nominal 80% prediction intervals.

	N1: random walk				N2: GARCH			
	Coverage		Relative width		Coverage		Relative width	
Model	$h = 1$	$h = 16$	$h = 1$	$h = 16$	$h = 1$	$h = 16$	$h = 1$	$h = 16$
Chronos-T5	0.73	0.76	0.95	0.94	0.73	0.77	0.83	0.84
Chronos-Bolt	0.83	0.76	1.18	0.97	0.84	0.77	1.10	0.89
Chronos-2	0.85	0.78	1.18	0.96	**0.80**	0.77	0.96	0.84
TiRex	0.77	**0.81**	0.97	1.05	0.77	0.81	0.87	0.93
Moirai-1.1	0.79	0.76	1.35	1.02	0.79	0.76	1.23	0.92
Moirai-2.0	0.79	0.74	1.05	0.96	0.78	0.77	0.92	0.85
TimesFM-2.0	**0.80**	0.84	1.07	1.12	**0.80**	**0.80**	0.99	0.94
TimesFM-2.5	0.87	0.84	1.22	1.08	0.88	**0.80**	1.17	0.93
Sundial	0.25	0.43	0.27	0.50	0.26	0.47	0.25	0.47
FinCast	0.78	0.83	1.13	1.11	0.78	0.82	1.04	1.01

The interval is the 0.1–0.9 band. Coverage is the fraction of futures inside it; relative width divides its mean width by $2.56\sigma\sqrt{h}$, the Gaussian N1 reference (only a scale reference on N2). Each rung uses 512 series and three seeds. Chronos-T5, Moirai-1.1 and Sundial use quantiles of 100 sampled paths; other rows use quantile heads. Time-MoE has no interval output in the evaluated interface. Bold marks the coverage closest to the nominal 0.80 in each column; widths are not ranked.

D Directional Bias and Input Controls

D.1 Direction and Recent-Slope Response

Figure 4 shows the signed mean and the upward fraction over the horizon, and Table 8 reports them at h=128 together with the correlation with the recent context slope; the series start at level $+100$.

Magnitude, sign and recent-slope response. Chronos-T5's upward fractions on N1 are 62, 79 and 85 percent at $h = 1$, 16 and 128, with mean $+3.4\sigma$ at $h = 128$ (109 of 128; two-sided exact binomial $p = 1.6 \times 10^{-16}$ against one half). Negating the same input series gives 82 percent upward and mean $+6.4\sigma$; mirror-pair sums of departures are $+9.8\sigma$ for this pair and $+3.9\sigma$ for the centred pair. Recent-slope correlations of 0.2–0.4 explain 4–18 percent of departure variance.

Chronos-Bolt has mean $+2.25\sigma$ and 79 percent upward forecasts at $h = 128$. Chronos-2 has $+0.61\sigma$, 70 percent upward and slope correlation $+0.42$. TiRex has $+2.88\sigma$, 85 percent upward and slope correlation -0.11, so its upward bias does not come from following the slope. Moirai-2.0 has $+0.08\sigma$ and 59 percent upward. TimesFM-2.5 has $+0.11\sigma$, 52 percent upward and slope correlation -0.42; disabling its built-in flip averaging changes the first two values to $+0.18\sigma$ and 52 percent. Sundial has $+0.46\sigma$, 52 percent upward and slope correlation -0.56. Its rms departure is 9.2σ, compared with Time-MoE's 11.1σ.

Uncertainty of the upward fractions. The 128 N1 series are independent, so each model's fraction of upward forecasts at h=128 can be tested against one half exactly. Two-sided exact binomial tests give $p = 1.6 \times 10^{-16}$ for Chronos-T5 and TiRex (109 of 128 each), 3×10^{-11} for Chronos-Bolt, 3×10^{-7} for Moirai-1.1 and 5×10^{-6} for Chronos-2; after a Holm correction over the eleven models these are 1.7×10^{-15} to 3×10^{-5}. Moirai-2.0 (76 of 128, $p = 0.04$) and FinCast (51 of 128, $p = 0.03$) are the only other models below 0.05 before correction, and neither is after it (0.21 and 0.16); the remaining four have $p \geq 0.33$. On the randomised daily windows the copies of one window are not independent, and a cluster bootstrap over family and end date gives standard errors of 0.008 to 0.015 for the same fractions; there Time-MoE and Sundial forecast a rise on 58 and 59 percent of copies, as often as Chronos-2 (59 percent), and FinCast a fall on 56 percent, so with real magnitudes the split by design is less clean than on the synthetic ladder.

Comparison across model families. Moirai-1.1 at its automatic patch setting has $+3.0\sigma$, 73 percent upward and slope correlation $+0.24$. TimesFM-2.0 has -0.03σ, 49 percent upward and slope correlation -0.34. Time-MoE has -0.47σ, 55 percent upward and slope correlation -0.62; on N2 and N4 its means are -1.16 and -2.01σ, with upward fractions 55 and 52 percent.

Table 8: Forecast direction under level, mirror and interface changes.

Model	Input or interface setting	h = 16		h = 128	
		Mean / σ	Upward fraction	Mean / σ	Upward fraction
Chronos-T5	Level +100	+0.55	0.79	+3.42	0.85
	Level +100, mirrored	+1.26	0.70	+6.37	0.82
	Zero-centred	+0.43	0.60	+2.07	0.60
	Zero-centred, mirrored	+0.13	0.52	+1.82	0.59
Chronos-Bolt	Default	+0.72	0.80	+2.25	0.79
Chronos-2	Default	+0.08	0.60	+0.61	0.70
TiRex	Default	+0.57	0.81	+2.88	0.85
Moirai-1.1	Patch auto (default)	+0.37	0.62	+2.96	0.73
	Patch 8 (yearly/quarterly)	+1.22	0.82	+4.79	0.84
	Patch 16 (weekly/daily)	+0.78	0.74	+3.52	0.90
	Patch 32 (daily/hourly)	+0.55	0.73	+3.00	0.88
	Patch 64 (hourly/minute)	–0.17	0.45	+0.06	0.48
	Patch 128 (minute/second)	–0.09	0.48	–0.11	0.46
Moirai-2.0	Default	–0.06	0.55	+0.08	0.59
TimesFM-2.0	Default (frequency 0)	+0.13	0.55	–0.03	0.49
	Declared daily or faster	+0.00	0.52	–0.09	0.52
	Declared weekly/monthly	–0.04	0.52	–0.29	0.47
	Declared quarterly/yearly	+0.03	0.45	+0.06	0.49
TimesFM-2.5	Default	+0.07	0.51	+0.11	0.52
	Flip averaging off	+0.09	0.54	+0.18	0.52
Time-MoE	Default	–0.52	0.53	–0.47	0.55
Sundial	Default	+0.10	0.54	+0.46	0.52
FinCast	Default	+0.01	0.51	–0.53	0.40
	Declared daily or faster	–0.12	0.45	–0.47	0.38
	Declared weekly/monthly	+0.89	0.64	+5.05	0.73
	Declared yearly	+0.53	0.57	+3.61	0.70

Each row uses 128 N1 series, context 512 and horizon 128. Mean is signed departure in σ units; upward fraction counts forecasts above the last value. The default, level and flip rows share one set of 128 series and the fixed-patch and declared-frequency sweeps another, so differences between a default row and a sweep row are not paired effects. Chronos-T5 varies offset or sign with shared increments. Moirai-1.1 labels give the frequencies associated with each patch size. Defaults follow Table 2; TimesFM-2.5 enables flip averaging unless stated otherwise.

The tested encoder-based and recurrent models show a consistent upward tendency, while the general-corpus decoder-only models have smaller or less consistent directional means. The pattern is observed across related checkpoints, whose corpora, laboratories, objectives and normalisation differ together, so it is reported descriptively; Section 4 varies the corpus alone.

D.2 Frequency Indicators and Patch Size

Interventions. TimesFM and FinCast take a three-level frequency indicator whose first level groups all frequencies up to daily; Moirai has five patch sizes associated with training frequency bands (Woo et al., 2024). We hold the values fixed and change only the declaration; for Moirai the input representation also changes with patch size. The other eight models take no declaration.

FinCast's indicator. With the daily declaration, FinCast has 38 percent upward forecasts and mean -0.5σ at $h = 128$, with negative recent-slope correlation. With the identical values declared weekly or yearly, these become 73 and 70 percent, with means $+5.1$ and $+3.6\sigma$. Pairwise rms differences over the horizon are 6–7σ, so **the indicator changes both the magnitude and the sign of the response on the null**.

Moirai's patch size and TimesFM's indicator. Moirai patches 8–32 give means $+3.0$–$+4.8\sigma$ and 84–90 percent upward forecasts. Patches 64 and 128 give $+0.06$ and -0.11σ, with 48 and 46 percent upward. Pairwise rms differences are 2.4–5.3σ. TimesFM's indicators produce rms

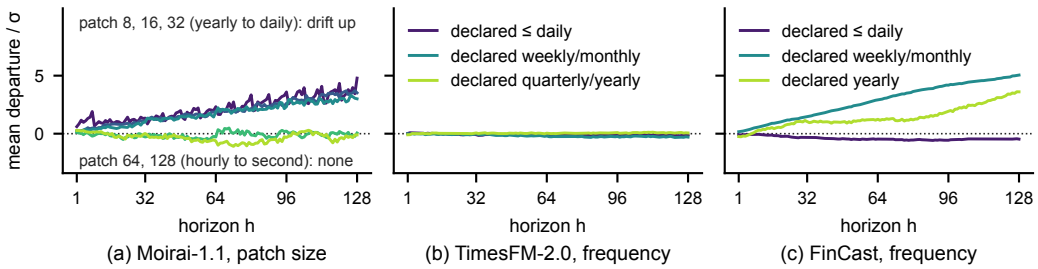


Figure 8: The same 128 N1 contexts under different frequency-associated settings: (a) Moirai patch size; (b) TimesFM indicator; (c) FinCast indicator. The dotted line marks the null's optimal signed departure, zero. Moirai's patch intervention also changes input representation.

differences of 1.8σ between the daily setting and the weekly or yearly settings, with upward fractions 52, 47 and 49 percent: its forecasts change without a common direction.

D.3 Spectral Shape and Forecast-Window Length

Frequency-band diagnostic. We compute the spectrum and spectral flatness of the departure $\hat{y} - y_t$, separating low-frequency power from seasonal bands, which are compared with a local continuum.

Low-frequency power and window sensitivity. On Chronos, periods longer than a quarter of the forecast window contain 82–86 percent of power on N1, N2 and N4. At $H = 128$, the period-24 band is 1.9–2.9 times its local continuum; at $H = 256$ this falls to 0.5–0.7. The largest apparent short-window peak is on N4, which has no periodic component. At $H = 256$, no period-7, 12 or 24 band exceeds its continuum by more than 1.24 times, so the seasonal peaks do not persist across window lengths.

Chronos has spectral flatness about 0.09, compared with Moirai's 0.45–0.56, TimesFM's 0.08 and Time-MoE's 0.03. Moirai's departure has relatively more high-frequency variation. The other three are dominated by smooth low-frequency components.

D.4 Raw Level and Input Normalisation

Level shifts have little effect on the ten models that centre their inputs, but change Chronos-T5's encoded input and token resolution. The two sweeps distinguish these effects.

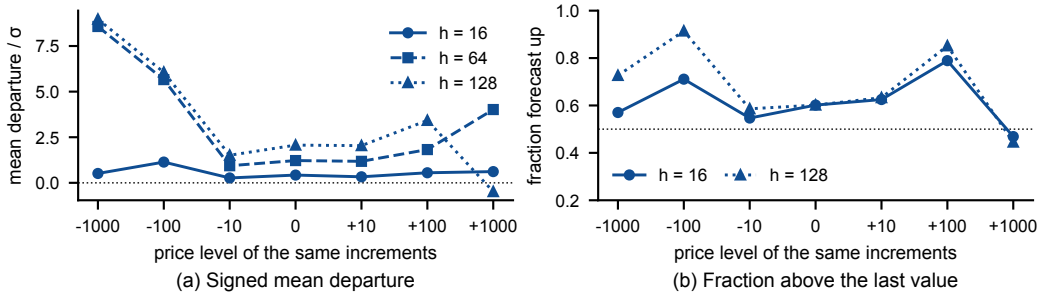


Figure 9: Chronos-T5 (small) on one set of N1 increments placed at seven price levels from -1000 to $+1000$ (128 series, H=128). (a) mean signed departure in units of σ at three horizons; (b) fraction of forecasts above the last value. The step is one percent of 100 at every level. At ± 1000 the step is approximately a tenth of a grid cell; the sweep changes both the scaled level and token resolution.

Six of the seven levels have a positive mean departure (Figure 9). At $+1000$, the exception, quantisation maps the context to a constant token sequence near scaled $+1$, and 45 percent of forecasts are

upward. Resolution is also coarse at -1000, where a grid cell is about 7σ wide, so the two endpoints mix the increments with their quantised representation.

Holding shape fixed across levels. Table 9 places the same 128 paths at levels -100, -10, 0, $+10$ and $+100$. Ten models centre and scale their inputs, so their normalised contexts are unchanged apart from numerical precision. Their departures generally agree to $10^{-4}\sigma$; the largest differences are $4\times10^{-3}\sigma$ for TimesFM-2.0, $2\times10^{-2}\sigma$ for FinCast, and 0.1 and 0.2σ for Time-MoE and TiRex. The latter two have root-mean-square differences of 0.01 and 0.04σ at h=128. Raw level therefore has little effect in these runs, including for the models with positive mean departures.

Chronos-T5 divides by mean absolute value without centring and then quantises, so a level shift changes its encoded input. Mean departures are $+1.4\sigma$ at -10, $+1.7$ at 0, $+2.1$ at $+10$, $+3.2$ at $+100$ and $+6.1$ at -100, with 56–86 percent of forecasts upward. The persistence of the sign at level zero rules out raw positive level as a necessary condition for this result.

Table 9: Sensitivity to a common shift in input level.

Model	Mean departure / σ					Upward fraction					Max. $\lvert\Delta D\rvert$
	−100	−10	0	+10	+100	−100	−10	0	+10	+100	vs. +100
Chronos-T5	+6.08	+1.43	+1.73	+2.07	+3.18	0.86	0.57	0.56	0.62	0.84	4.4×10^{1}
Chronos-Bolt	+2.25	+2.25	+2.25	+2.25	+2.25	0.79	0.79	0.79	0.79	0.79	5.3×10^{-5}
Chronos-2	+0.61	+0.61	+0.61	+0.61	+0.61	0.70	0.70	0.70	0.70	0.70	4.0×10^{-5}
TiRex	+2.88	+2.89	+2.88	+2.88	+2.88	0.85	0.85	0.85	0.85	0.85	2.1×10^{-1}
Moirai-1.1	+2.96	+2.96	+2.96	+2.96	+2.96	0.73	0.73	0.73	0.73	0.73	1.1×10^{-4}
Moirai-2.0	+0.08	+0.08	+0.08	+0.08	+0.08	0.59	0.59	0.59	0.59	0.59	1.8×10^{-4}
TimesFM-2.0	−0.03	−0.03	−0.03	−0.03	−0.03	0.49	0.49	0.49	0.49	0.49	4.3×10^{-3}
TimesFM-2.5	+0.11	+0.11	+0.11	+0.11	+0.11	0.52	0.52	0.52	0.52	0.52	7.6×10^{-5}
Time-MoE	−0.47	−0.47	−0.47	−0.47	−0.47	0.55	0.55	0.55	0.55	0.55	1.0×10^{-1}
Sundial	+0.47	+0.47	+0.47	+0.47	+0.47	0.55	0.55	0.55	0.55	0.55	1.1×10^{-4}
FinCast	−0.53	−0.53	−0.53	−0.53	−0.53	0.40	0.40	0.40	0.40	0.40	1.9×10^{-2}

The same 128 N1 paths ($\sigma = 1$) are placed at the five column levels; mean and upward fraction are evaluated at $h = 128$. The final column is the largest absolute change in departure relative to level +100, over all tested levels, series and forecast steps. It measures individual forecast changes, not a difference of the displayed means. Chronos-T5 samples its forecasts, and its row comes from a separate run on the same paths from that of Figure 9, so the two differ by sampling noise (for example $+3.18$ against $+3.42$ at level $+100$); Sundial, which also samples, differs from Table 8 in the same way ($+0.47$ and 0.55 against $+0.46$ and 0.52).

D.5 Chronos-T5 Model Size

Increasing Chronos-T5 size improves its N4 point estimates without a monotone reduction in N1 risk. The sweep uses identical contexts across checkpoints.

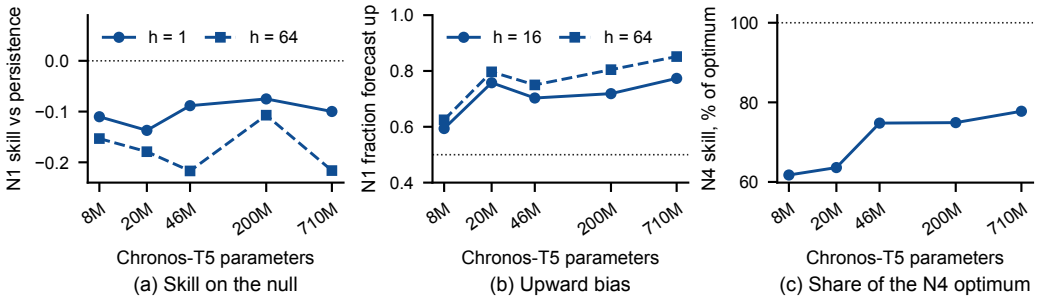


Figure 10: Five sizes of Chronos-T5 on the same N1 and N4 contexts (128 series, H=64): (a) N1 skill against persistence; (b) fraction of N1 forecasts above the last value; (c) N4 skill at h=1 as a share of the optimum, the conditional mean of Appendix B.1.

On N4, the fraction of the conditional-mean optimum's gain captured by Chronos-T5 rises from 62 percent at 8M parameters to 78 percent at 710M (Figure 10). On N1, skill at h=1 ranges from -7.5 to -14 percent without a monotone trend. The largest model forecasts upward on 85 percent of contexts, compared with 62 percent for the smallest; correlation with the recent slope falls from 0.57 to 0.38. Within this family, better N4 accuracy does not bring lower null risk.

Table 10: Context length, forecast direction and null skill.

Model	Metric	Context length T 128	256	512	1024	2048
Chronos-T5	Mean / σ, h = 128	+4.79 ± 0.37	+3.99 ± 0.31	+3.42 ± 0.36	NE	NE
	Upward fraction, h = 128	0.92	0.91	0.85	NE	NE
	Skill, h = 16	**+0.039**	−0.006	+0.003	NE	NE
Chronos-Bolt	Mean / σ, h = 128	+1.19 ± 0.26	+1.66 ± 0.42	+2.05 ± 0.31	+2.01 ± 0.25	+0.48 ± 0.28
	Upward fraction, h = 128	0.73	0.60	0.70	0.72	0.58
	Skill, h = 16	−0.136	−0.175	−0.141	−0.110	**−0.068**
Chronos-2	Mean / σ, h = 128	−0.13 ± 0.39	+0.13 ± 0.39	+0.61 ± 0.36	+0.53 ± 0.24	+0.28 ± 0.16
	Upward fraction, h = 128	0.55	0.70	0.70	0.71	0.63
	Skill, h = 16	−0.102	−0.013	−0.014	**+0.009**	−0.023
TiRex	Mean / σ, h = 128	+2.09 ± 0.24	+2.09 ± 0.29	+2.83 ± 0.29	+2.94 ± 0.28	+2.02 ± 0.29
	Upward fraction, h = 128	0.84	0.81	0.84	0.84	0.77
	Skill, h = 16	−0.135	**−0.083**	−0.132	−0.120	−0.154
Moirai-1.1	Mean / σ, h = 128	+5.09 ± 0.38	+3.61 ± 0.50	+2.31 ± 0.46	+1.90 ± 0.40	+1.63 ± 0.37
	Upward fraction, h = 128	0.92	0.77	0.66	0.69	0.66
	Skill, h = 16	−0.202	**−0.148**	−0.197	−0.291	−0.737
Moirai-2.0	Mean / σ, h = 128	+0.64 ± 0.29	+0.54 ± 0.34	+0.76 ± 0.44	+0.71 ± 0.37	+0.19 ± 0.28
	Upward fraction, h = 128	0.57	0.57	0.58	0.59	0.50
	Skill, h = 16	−0.080	**−0.066**	−0.134	−0.089	−0.104
TimesFM-2.0	Mean / σ, h = 128	+0.32 ± 0.55	+0.43 ± 0.37	+0.02 ± 0.27	+0.30 ± 0.22	+0.01 ± 0.20
	Upward fraction, h = 128	0.57	0.54	0.53	0.52	0.50
	Skill, h = 16	−0.485	−0.310	−0.221	−0.133	**−0.043**
TimesFM-2.5	Mean / σ, h = 128	+0.17 ± 0.26	+0.16 ± 0.25	+0.14 ± 0.15	+0.20 ± 0.16	+0.10 ± 0.11
	Upward fraction, h = 128	0.57	0.53	0.57	0.52	0.53
	Skill, h = 16	−0.145	−0.117	−0.055	−0.058	**−0.016**
Time-MoE	Mean / σ, h = 128	+0.18 ± 0.61	+0.96 ± 0.76	+0.47 ± 0.98	−0.06 ± 1.42	−2.13 ± 1.89
	Upward fraction, h = 128	0.55	0.54	0.58	0.51	0.49
	Skill, h = 16	**−0.803**	−0.906	−0.924	−2.791	−6.141
Sundial	Mean / σ, h = 128	+0.40 ± 0.57	+0.93 ± 0.79	+1.42 ± 0.75	+1.18 ± 0.82	+0.25 ± 1.04
	Upward fraction, h = 128	0.53	0.53	0.55	0.55	0.54
	Skill, h = 16	−0.685	−0.570	−0.474	−0.430	**−0.384**
FinCast	Mean / σ, h = 128	−0.60 ± 0.19	−0.47 ± 0.15	−0.53 ± 0.17	NE	NE
	Upward fraction, h = 128	0.33	0.40	0.40	NE	NE
	Skill, h = 16	−0.097	**−0.034**	−0.061	NE	NE

Each model uses 128 N1 paths, with nested contexts and fixed 128-step futures: Chronos-T5 and FinCast the direction contexts of Section 3, the other models 2048-point paths (Table 3). For the sampled means of Chronos-T5, Moirai-1.1 and Sundial, skill here also removes the Monte Carlo variance from the squared error. ±: standard error over series. NE: not evaluated beyond the 512-point interface cap (Chronos-T5, FinCast). Bold marks the highest skill within each model.

D.6 Context Length and Residual Bias

Table 10 uses 128 N1 paths per model with fixed futures and nested contexts from 128 to 2048 points (512 at most for Chronos-T5 and FinCast). At h=128 the mean departure falls from $+4.8$ to $+3.4\sigma$ for Chronos-T5, from $+5.1$ to $+1.6\sigma$ for Moirai-1.1 and from $+2.0$ to $+0.5\sigma$ for Chronos-Bolt, while TiRex stays between $+2.0$ and $+2.9\sigma$; at the longest contexts the five models that are not decoder-only still forecast a rise on 58 to 85 percent of contexts. Null skill does not improve uniformly: at h=16 TimesFM-2.0 improves from -0.49 to -0.04 while Time-MoE falls from -0.80 to -6.14, and Sundial improves at h=16 but declines at h=128 (-0.45 to -1.41).

E Corpus Intervention Protocol and Seed Variation

E.1 Corpus Intervention and Training Design

The synthetic corpus mixes seasonal AR(1) series (weight 0.45), the same with a linear trend (0.25), random walks with drift plus seasonality (0.20) and N4-type bounce series with a random spread (0.10), in units of a unit innovation. In the *growth* corpus the trends and drifts have mean $+0.02$ per step, so that its trend components favour rising series; in the *symmetric* corpus the means are zero.

Two models are trained on each corpus. The masked encoder uses 32 context patches of 16 points and 8 future mask tokens with bidirectional attention, following the design of Moirai. It decodes the future patches jointly and applies loss only to them. The decoder-only model follows TimesFM and

Table 11: **A corpus with positive trends makes both model designs forecast a rise.**

		Held-out MSE		N1, h = 128			N1 skill	N4 capture
Model design	**Corpus**	**Model**	**Persistence**	**Mean / σ**	**Upward**	ρ_{even}	h = 16	h = 1
Masked encoder	Growth	**0.700**	1.541	+1.49 ± 0.06	0.86	0.35	–0.042	0.38
	Symmetric	**0.731**	1.576	+0.14 ± 0.04	0.53	0.03	–0.040	0.38
Decoder-only	Growth	**0.714**	1.541	+1.84 ± 0.30	0.83	0.33	–0.025	0.46
	Symmetric	**0.749**	1.576	+0.79 ± 0.66	0.65	0.14	–0.018	0.44

Entries average three training seeds; ± denotes their standard deviation for mean departure. Held-out MSE uses 4096 corpus windows in normalised units; bold marks the lower MSE within each row. N1 direction uses 128 series; ρ_{even} is the even-energy ratio. N4 capture uses the exact conditional-mean oracle. The designs also differ in loss positions and decoding.

Time-MoE in using causal attention and next-patch prediction at every position, with autoregressive inference. The designs thus differ in attention, loss positions and decoding.

Both normalise the context by its own mean and standard deviation, share four pre-LN transformer layers (d=128, four heads, 0.54M parameters) and a linear patch head, and are trained with the squared error in normalised units for 12000 steps of 256 windows with AdamW, three training seeds each.

Evaluation uses the released models' direction contexts from Section 3 (N1, 128 series, H=128) and their mirrors, the ladder's N1 and N4 at H=16, and 4096 held-out corpus windows to assess fit.

E.2 Seed-Level Direction Checks

Under the symmetric corpus, two decoder seeds have mean N1 departures more than twice their per-seed standard errors above zero, while none is below; the corresponding encoder counts are zero. The aggregate means and between-seed standard deviations appear in Table 11.

E.3 Larger Models

We repeated the intervention with larger models, stating one prediction before the runs: the growth corpus produces an upward N1 bias in both designs. The designs, corpora, loss, schedule and evaluation are those above; only the width and depth change, to d=256 and six layers (3.18M parameters, about six times as many), with three seeds per design and corpus. **The prediction holds in every seed.** On the growth corpus the masked encoder's mean N1 departure at h=128 is $+1.39 \pm 0.14\sigma$ with 82 percent of forecasts upward, and the decoder-only model's $+1.93 \pm 0.39\sigma$ with 84 percent, against $+1.49$ and $+1.84$ at 0.54M parameters (Table 12); every seed of both designs lies between $+1.23$ and $+2.38\sigma$. On the symmetric corpus the encoder again stays near balance ($+0.16 \pm 0.12\sigma$, 51 percent upward), and the decoder again depends on the seed, from $+0.19$ to $+1.30\sigma$ ($+0.62 \pm 0.60$ on average). Both designs beat persistence on held-out corpus windows. Six times as many parameters thus leave the corpus effect, and the decoder's seed-dependent residual bias, as they were.

Table 12: Training-corpus effects within two model designs with 3.18M parameters.

		Held-out MSE		N1, h = 128			N1 skill	N4 capture
Model design	**Corpus**	**Model**	**Persistence**	**Mean / σ**	**Upward**	ρ_{even}	h = 16	h = 1
Masked encoder	Growth	**0.700**	1.541	+1.39 ± 0.14	0.82	0.30	–0.041	0.34
	Symmetric	**0.732**	1.576	+0.16 ± 0.12	0.51	0.04	–0.040	0.36
Decoder-only	Growth	**0.694**	1.541	+1.93 ± 0.39	0.84	0.41	–0.017	0.64
	Symmetric	**0.729**	1.576	+0.62 ± 0.60	0.60	0.16	–0.013	0.65

As Table 11, with d=256 and six layers instead of d=128 and four. Entries average three training seeds; ± denotes their standard deviation for mean departure. Bold marks a held-out MSE below persistence's.

E.4 Trend Statistics of Public Pretraining Corpora

The controlled experiment shows that a corpus with positive trends is enough to produce an upward bias. To see whether public pretraining data have such trends, we computed descriptive statistics for three corpora: the datasets released with Chronos (26 subsets), LOTSA, the pretraining archive of Moirai-1.1 (91 subsets), and Time-300B, the pretraining corpus of Time-MoE (97 subsets). From each subset we drew up to 5000 series and four end points per series, with 64 to 512 context points, and measured the change over the next h steps in units of the standard deviation of the context's one-step changes, the unit of the N1 probe. We made no prediction for this audit.

Table 13 reports h=128, for the 26, 71 and 74 subsets with series long enough. In the typical subset the series do not trend: the median change is zero or nearly so and half of the moves are rises. Pooled over series, the Chronos datasets and LOTSA lean upward, with 57 percent of moves rising and a mean change of about $+1.4\sigma$ once each change is clipped to $\pm 10\sigma$; Time-300B leans less (52 percent, $+0.4\sigma$). Almost all of the lean comes from the M4 subsets, in which 72 to 91 percent of monthly, quarterly and yearly moves are rises; without them the pooled share is 0.49 to 0.50 and the clipped mean -0.01 to $+0.13\sigma$. Unclipped means are far larger, $+19$ to $+23\sigma$, because a few series with nearly constant contexts make very large moves. A widely used source of pretraining series thus has the kind of trend that produces the bias in the controlled experiment; how each model weights these subsets is not public.

Table 13: Trend statistics of three public pretraining corpora at h=128.

			Share of moves that rise			Subsets leaning		Change (σ)	
Corpus	**Subsets**	**Series**	**Median**	**Pooled**	**No M4**	**Up**	**Down**	**Median**	**Clipped mean**
Chronos datasets	26	66 105	0.51	0.57	0.49	0.23	0.04	+0.00	+1.44 (–0.01)
LOTSA (Moirai)	71	64 527	0.50	0.57	0.50	0.20	0.15	+0.00	+1.46 (+0.13)
Time-300B (Time-MoE)	74	118 798	0.51	0.52	0.50	0.28	0.09	+0.02	+0.43 (+0.05)

Up to 5000 series per subset and four windows per series; only subsets with series long enough for h=128. Change is $(x_{t+128} - x_t)$ divided by the standard deviation of the context's one-step changes. Share of moves that rise: rises over rises and falls, ignoring zero changes. Median: over subsets; pooled: over series; No M4: pooled without the M4 subsets. Subsets leaning: share of subsets whose rising share exceeds 0.55 (up) or is below 0.45 (down). Clipped mean: pooled mean of the change clipped to $\pm 10\sigma$, without M4 in parentheses.

F Observed Data and Predictable Alternatives

F.1 Daily Financial Windows and Sign-Randomised Nulls

Figure 6 shows the copies and the raw windows; in its panel (c) Time-MoE lies below the axis, between -0.9 and -1.7 from h=16 on, and its model curves are those of Table 1.

Sources and windows. The two public sources were downloaded on 2026-09-16. Exchange rates are the European Central Bank's daily euro reference rates, from which twenty floating currencies are used (USD, JPY, GBP, CHF, AUD, CAD, SEK, NOK, NZD, MXN, ZAR, BRL, KRW, PLN, HUF, CZK, SGD, INR, ILS, IDR; pegged and managed currencies are not); the rate itself is the level series.

Equities are the Kenneth R. French data library's daily value-weighted returns of 49 industry portfolios and the market return (Mkt-RF plus RF), all built from CRSP, with levels rebuilt as $100 \prod(1+r)$; every portfolio with a complete record since 2015 is used.

Observations before 2015-01-01 are ignored. Per series, windows of $512 + 128$ trading days are taken at a stride of 128 counting back from the last observation, so that the 128-day futures do not overlap while consecutive contexts do, giving 1280 windows (380 exchange-rate, 900 equity) whose futures end between 2017-09-12 and 2026-09-15; 280 of them end after 2025-01-01, past the published training cut-offs of Chronos, Moirai, TimesFM-2.0 and Time-MoE. The median daily return standard deviation σ_r over a context is 0.46 percent for the exchange rates and 1.41 percent for the portfolios.

An exact null from observed return magnitudes. For a window $y_0, \ldots, y_{639}$ with simple returns $r_t = y_t/y_{t-1} - 1$ draw $\varepsilon_t \in \{-1, +1\}$ i.i.d. with equal probability, independent of everything, and

set $\tilde{y}_0 = y_0$, $\tilde{y}_t = \tilde{y}_{t-1}(1 + \varepsilon_t r_t)$. Since ε_t is independent of $\tilde{\mathcal{F}}_{t-1}$ and of r_t and has mean zero, $\mathbb{E}[\tilde{y}_t \mid \tilde{\mathcal{F}}_{t-1}] = \tilde{y}_{t-1}$: the surrogate satisfies Definition 1 exactly, while $|\tilde{r}_t| = |r_t|$, so the real volatility clustering, heavy tails and calendar gaps are kept.

Each window gets $K{=}4$ independent sign draws, pooled for the null evaluation and used for the sign-averaged correction. The realised move of the surrogates at $h{=}128$ is $+0.07\sigma_r$ with 47 percent of copies up (compounded-return distributions need not have equal means and medians), against $+3.5\sigma_r$ and 65 percent for the raw windows.

Return-unit metrics and clustered uncertainty. We report departure $(\hat{y}_h/y_T - 1)/\sigma_r$, where σ_r is the standard deviation of the context's simple returns, the fraction of windows with $\hat{y}_h > y_T$, and skill $1 - \mathrm{MSE}/\mathrm{MSE}_{\text{persistence}}$. The errors for skill are $((y - \hat{y})/y_T)^2$, pooled over windows in return units. This removes differences in price-level scale; the exchange rates span four orders of magnitude. Unlike the departure diagnostic, this skill ratio does not divide each window's squared error by σ_r^2. Higher-volatility windows can therefore contribute more to it. Scaling each window's squared errors by its σ_r^2, as for the size deciles (Appendix F.4), weights calm and volatile windows alike. On the 900 equity windows at $h{=}128$ the mirror correction then lowers skill for nine models instead of ten, by more than two standard errors for the same four (TiRex, Time-MoE, Sundial and FinCast); Moirai-1.1's change becomes $+0.009 \pm 0.073$. The equity windows form 18 of the 37 clusters. For Chronos and Moirai the Monte Carlo term is removed from the root-mean-square departure as in Appendix C.2.

Standard errors come from a cluster bootstrap over calendar windows (clusters are family $\times$ future end date, 37 in all; a window's surrogates share its cluster), because the 50 equity portfolios share calendar windows. Readouts follow Appendix B.2; TimesFM-2.0 and FinCast use the daily declaration.

Departures on the randomised and raw windows. On the copies the signs of the departures hold within the FX and equity splits (Table 14). TimesFM-2.0 and Time-MoE have positive return-unit means of $+0.48 \pm 0.12$ and $+1.96 \pm 0.27\sigma_r$ despite negative correlations with the recent slope (-0.18 and -0.54), and level-unit means of $+0.21$ and $+0.38\sigma$. Moirai-2.0 and TimesFM-2.5 are close to balanced (49 and 51 percent upward) yet lose at both horizons; Sundial has $+1.74\sigma_r$ (59 percent upward) with slope correlation -0.53. At $h{=}16$, Chronos-T5, Moirai-1.1, TimesFM-2.0 and FinCast lose 5–21 percent and Time-MoE 123 percent.

On the raw windows, mean departures are $+2.17{\pm}0.21\sigma_r$ for Chronos-T5, $+2.38{\pm}0.24$ for Moirai-1.1 (the copy-to-raw ratios are 71 and 96 percent), $+1.41$ for Moirai-2.0 (67 percent upward), -3.16 for Sundial (29 percent upward) and -4.33 ± 0.93 for Time-MoE. Chronos-T5's $h{=}128$ skill is $+0.05 \pm 0.06$ on equities, whose mean move was $4.3\sigma_r$, -0.17 ± 0.10 on FX and $+0.07 \pm 0.05$ after 2025; Chronos-Bolt and TiRex have pooled raw skill near $+0.12$ and post-2025 point estimates near zero; Time-MoE loses 89 percent, and FinCast is at -0.00 ± 0.03 pooled and -0.16 ± 0.03 after 2025 (Tables 23 and 24).

Resolution and pretraining overlap. At exchange-rate scale, σ_r of 0.46 percent, Chronos's 4096-token grid has a spacing of about $1.6\sigma_r$, coarser than one step, so its $h{=}1$ departures on these windows are dominated by quantisation and the anchor is read from $h{=}16$ on; the equity windows, at 1.41 percent, sit near the synthetic ladder's scale.

Public daily exchange rates and equity indices can appear in pretraining corpora (Chronos and Moirai list an exchange-rate dataset that ends in 2016, Time-300B has a finance domain, FinCast is finance-native). The copies are new sequences by construction, and the post-2025 split of Table 14 postdates the published training cut-offs of the four 2024 general-corpus models. On pre-2015 equity windows several models appear to remember the market, and M4 daily holds its path up to October 2017; on this anchor the timing and calendar-shift checks find none, and M4 daily covers part of the future of only its earliest equity end date (Appendix F.5).

F.2 Nasdaq Transaction Prices

Transaction sampling and the MA(1) baseline. This anchor tests whether N4's bid–ask-bounce comparison carries over to observed trades. It uses one day of Nasdaq's public TotalView-ITCH

Table 14: Daily forecast departures and skill by data subset.

Model	Sign-randomised		Raw daily data		
	Mean / σ_r $h = 128$	Skill $h = 16$	Mean / σ_r $h = 128$	Skill $h = 16$	Skill $h = 128$
FX (380 windows)					
Chronos-T5	+1.70 ± 0.19	−0.092 ± 0.056	+1.06 ± 0.21	−0.181 ± 0.158	−0.172 ± 0.100
Chronos-Bolt	+1.65 ± 0.08	−0.042 ± 0.021	+1.71 ± 0.25	**−0.067** ± 0.053	**−0.091** ± 0.084
Chronos-2	+0.39 ± 0.09	**−0.019** ± 0.016	−0.03 ± 0.22	−0.198 ± 0.129	−0.127 ± 0.045
TiRex	+2.11 ± 0.11	−0.070 ± 0.016	+0.76 ± 0.56	−0.142 ± 0.056	−0.134 ± 0.103
Moirai-1.1	+2.18 ± 0.15	−0.156 ± 0.069	+1.75 ± 0.24	−0.259 ± 0.063	−0.230 ± 0.110
Moirai-2.0	−0.04 ± 0.12	−0.058 ± 0.020	−0.17 ± 0.54	−0.184 ± 0.055	−0.173 ± 0.073
TimesFM-2.0	+0.08 ± 0.14	−0.091 ± 0.022	−1.70 ± 0.64	−0.198 ± 0.099	−0.260 ± 0.099
TimesFM-2.5	+0.11 ± 0.09	−0.101 ± 0.022	−1.31 ± 0.42	−0.247 ± 0.100	−0.152 ± 0.058
Time-MoE	+0.80 ± 0.27	−0.987 ± 0.141	−2.45 ± 1.01	−1.545 ± 0.401	−0.809 ± 0.245
Sundial	+0.96 ± 0.14	−0.251 ± 0.042	−1.51 ± 0.85	−0.704 ± 0.264	−0.511 ± 0.184
FinCast	−0.40 ± 0.05	−0.058 ± 0.024	−1.00 ± 0.19	−0.070 ± 0.043	−0.101 ± 0.040
Equities (900 windows)					
Chronos-T5	+1.46 ± 0.08	−0.048 ± 0.010	+2.65 ± 0.25	−0.057 ± 0.045	+0.048 ± 0.058
Chronos-Bolt	+1.37 ± 0.13	−0.097 ± 0.028	+3.72 ± 0.42	−0.274 ± 0.121	**+0.129** ± 0.110
Chronos-2	+0.06 ± 0.08	**−0.038** ± 0.012	+1.93 ± 0.30	−0.045 ± 0.024	−0.020 ± 0.050
TiRex	+1.79 ± 0.11	−0.051 ± 0.011	+2.19 ± 0.40	−0.076 ± 0.058	+0.123 ± 0.100
Moirai-1.1	+2.32 ± 0.18	−0.210 ± 0.041	+2.64 ± 0.32	−0.119 ± 0.098	−0.016 ± 0.075
Moirai-2.0	−0.15 ± 0.16	−0.083 ± 0.014	+2.08 ± 0.50	−0.251 ± 0.105	−0.004 ± 0.092
TimesFM-2.0	+0.65 ± 0.14	−0.208 ± 0.071	+0.50 ± 0.82	−0.743 ± 0.639	−0.186 ± 0.187
TimesFM-2.5	+0.18 ± 0.10	−0.079 ± 0.013	+1.22 ± 0.36	−0.092 ± 0.047	−0.009 ± 0.060
Time-MoE	+2.45 ± 0.31	−1.244 ± 0.213	−5.13 ± 1.23	−1.182 ± 0.517	−0.891 ± 0.299
Sundial	+2.07 ± 0.26	−0.335 ± 0.063	−3.85 ± 0.78	−0.152 ± 0.136	−0.542 ± 0.200
FinCast	−0.20 ± 0.08	−0.065 ± 0.013	−0.07 ± 0.24	**−0.023** ± 0.040	−0.001 ± 0.030
Post-2025 subset (280 windows)					
Chronos-T5	+1.51 ± 0.15	−0.047 ± 0.016	+2.29 ± 0.35	−0.097 ± 0.095	**+0.074** ± 0.046
Chronos-Bolt	+1.80 ± 0.11	**−0.019** ± 0.029	+2.98 ± 0.52	−0.183 ± 0.086	−0.010 ± 0.096
Chronos-2	+0.11 ± 0.15	−0.039 ± 0.016	+1.70 ± 0.32	−0.013 ± 0.033	+0.070 ± 0.056
TiRex	+1.80 ± 0.09	−0.051 ± 0.020	+1.38 ± 0.50	−0.087 ± 0.052	+0.019 ± 0.052
Moirai-1.1	+2.19 ± 0.22	−0.093 ± 0.029	+2.58 ± 0.41	−0.052 ± 0.029	−0.133 ± 0.081
Moirai-2.0	−0.17 ± 0.23	−0.104 ± 0.016	+1.63 ± 0.36	−0.123 ± 0.085	+0.023 ± 0.035
TimesFM-2.0	+0.16 ± 0.15	−0.092 ± 0.037	+0.52 ± 0.33	−0.032 ± 0.075	−0.159 ± 0.047
TimesFM-2.5	−0.03 ± 0.08	−0.068 ± 0.014	−0.05 ± 0.30	−0.053 ± 0.037	−0.096 ± 0.052
Time-MoE	+1.19 ± 0.43	−0.807 ± 0.247	−6.88 ± 1.03	−0.843 ± 0.729	−1.433 ± 0.492
Sundial	+1.11 ± 0.26	−0.267 ± 0.077	−4.85 ± 0.71	−0.054 ± 0.178	−0.871 ± 0.207
FinCast	−0.38 ± 0.07	−0.096 ± 0.017	−1.29 ± 0.16	**+0.018** ± 0.089	−0.156 ± 0.033

Contexts 512, futures 128, four sign-randomised copies per window; means in σ_r units; skill pools squared errors in return units. ±: cluster-bootstrap SE over market and future end date (37 clusters; 19 FX, 18 equities, 8 post-2025). Mean realised moves at $h = 128$ (sign-randomised / raw), in σ_r: FX $-0.71/+1.64$; equities $+0.39/+4.33$; post-2025 $+0.29/+3.94$. Persistence skill is zero. Bold marks the highest model skill within each panel and column.

data (2019-12-30). Every execution is extracted, fills sharing a timestamp are collapsed to one transaction, and the 40 most actively traded symbols priced above one dollar are selected. Up to 25 blocks per symbol, each of $512+16$ consecutive transactions, are spaced evenly through the session, yielding 792 windows. Each has $K=4$ multiplicative sign-randomised copies. Metrics use return units; median σ_r is 1.1 basis points per transaction.

The transactions differ from Roll's model. Fifty-five percent of consecutive transactions print at the same price, the trade direction flips only 27 percent of the time where independent directions would flip half the time, the lag-one autocorrelation of returns is -0.01 on average and below -0.1 in a fifth of the windows.

An MA(1) fitted to each context has skill -0.05 ± 0.03 at h=1, -0.22 on the windows whose context autocorrelation is below -0.1: the fitted MA(1) does not use bounce successfully at this frequency in this sample. Only the sign-randomised copies have the constructed mean-null guarantee (Table 15).

Directional patterns and point risk. On the sign-randomised copies the signs of the daily anchor return at trade frequency: at h=16 Chronos-Bolt, TiRex and Moirai-1.1 forecast a rise in 67, 67 and 65 percent of windows, $+0.41$, $+0.31$ and $+0.44\sigma_r$; Chronos-2 in 53; the five general-corpus decoder-only models between 47 and 51; FinCast in 45.

On the raw windows four models sit above persistence at h=1, Moirai-2.0 at $+0.08 \pm 0.08$, Chronos-2 at $+0.05 \pm 0.10$, TimesFM-2.5 and Chronos-Bolt at $+0.02$, none by more than one standard error, and at h=16 every model is below it.

At h=16, the mirror increases skill on the raw windows for every model displayed in Table 15. It also increases skill on the sign-randomised copies except for TimesFM-2.5, which changes from -0.071 to -0.072. At h=1, the changes are mixed: Chronos-Bolt, TiRex, TimesFM-2.5 and Time-MoE lose skill on the raw windows, and Moirai-2.0 and TimesFM-2.5 lose skill on the copies. TimesFM-2.5's released additive correction does not make it exactly invariant to the multiplicative mirror used here.

Token resolution and single-day scope. Chronos-T5 is left out of the table: its token grid is 67 steps wide at the median window and wider than a step in every window (skill -15 at h=1, departure $+34\sigma_r$ on the copies), a resolution limit that the continuous-input Chronos-Bolt and Chronos-2 avoid. The anchor extends the directional comparison to transaction frequency; it does not provide a raw-data positive control, and one day does not sample variation across days or regimes.

Table 15: Raw and mirror-corrected forecasts on Nasdaq transaction windows.

(a) Raw transaction prices

Forecast	Skill, h = 1		Skill, h = 16	
	Raw	Mirror	Raw	Mirror
Persistence	0	NE	0	NE
MA(1), fitted	−0.051 ± 0.032	NE	−0.000 ± 0.007	NE
Chronos-Bolt	**+0.018** ± 0.115	+0.014 ± 0.095	−0.115 ± 0.050	**−0.060** ± 0.032
Chronos-2	+0.047 ± 0.105	**+0.067** ± 0.087	−0.041 ± 0.052	**−0.009** ± 0.048
TiRex	**−0.030** ± 0.040	−0.040 ± 0.047	−0.033 ± 0.044	**−0.021** ± 0.036
Moirai-1.1	−0.985 ± 0.466	**−0.476** ± 0.274	−0.128 ± 0.135	**−0.072** ± 0.072
Moirai-2.0	+0.082 ± 0.085	**+0.091** ± 0.070	−0.009 ± 0.057	**−0.003** ± 0.043
TimesFM-2.0	−0.115 ± 0.106	**+0.001** ± 0.095	−0.191 ± 0.070	**−0.097** ± 0.054
TimesFM-2.5	**+0.019** ± 0.059	+0.016 ± 0.061	−0.023 ± 0.041	**−0.021** ± 0.041
Time-MoE	**−0.095** ± 0.113	−0.107 ± 0.107	−0.530 ± 0.173	**−0.523** ± 0.166
Sundial	−0.064 ± 0.101	**−0.048** ± 0.090	−0.252 ± 0.137	**−0.222** ± 0.126
FinCast	−0.259 ± 0.141	**−0.197** ± 0.102	−0.071 ± 0.044	**−0.024** ± 0.035

(b) Sign-randomised copies

Model	Departure, h = 16		Skill, h = 1		Skill, h = 16	
	Mean / σ_r	Upward	Raw	Mirror	Raw	Mirror
Chronos-Bolt	+0.41 ± 0.02	0.67	−0.199 ± 0.065	**−0.136** ± 0.068	−0.102 ± 0.042	**−0.088** ± 0.031
Chronos-2	+0.04 ± 0.02	0.53	−0.103 ± 0.072	**−0.085** ± 0.046	−0.135 ± 0.033	**−0.122** ± 0.029
TiRex	+0.31 ± 0.02	0.67	−0.067 ± 0.034	**−0.046** ± 0.023	−0.129 ± 0.034	**−0.087** ± 0.026
Moirai-1.1	+0.44 ± 0.02	0.65	−1.284 ± 0.323	**−0.518** ± 0.124	−0.137 ± 0.046	**−0.048** ± 0.028
Moirai-2.0	−0.09 ± 0.02	0.47	**−0.036** ± 0.034	−0.039 ± 0.040	−0.118 ± 0.054	**−0.047** ± 0.036
TimesFM-2.0	−0.14 ± 0.04	0.47	−0.161 ± 0.051	**−0.097** ± 0.037	−0.207 ± 0.048	**−0.173** ± 0.043
TimesFM-2.5	+0.03 ± 0.02	0.51	**−0.057** ± 0.024	−0.059 ± 0.025	**−0.071** ± 0.027	−0.072 ± 0.028
Time-MoE	−0.08 ± 0.05	0.49	−0.134 ± 0.059	**−0.102** ± 0.058	−0.660 ± 0.152	**−0.605** ± 0.136
Sundial	+0.03 ± 0.03	0.51	−0.059 ± 0.032	**−0.045** ± 0.030	−0.219 ± 0.057	**−0.190** ± 0.055
FinCast	−0.06 ± 0.02	0.45	−0.263 ± 0.068	**−0.156** ± 0.059	−0.064 ± 0.029	**−0.019** ± 0.022

792 windows from 40 symbols on 2019-12-30; context 512 transactions, future 16, four sign copies per window. Skill uses pooled squared return errors (higher is better); mean departure uses σ_r units. ±: window-bootstrap SE. NE: correction not evaluated for the two baselines. The window bootstrap does not capture between-day variation. Chronos-T5 is excluded because its token grid cannot resolve typical increments. Bold marks the higher of raw and mirror skill within each row, horizon and panel.

F.3 ETTh1 Forecasts against Persistence

ETTh1 provides a structured nonfinancial comparison on which departures from persistence can improve point accuracy.

The hourly transformer-load and oil-temperature data are split into 512-point contexts and 128-point futures at stride 128: 924 windows across seven columns. Each model forecasts each window and its additive mirror. Skill is pooled in units of each context's one-step σ, with bootstrap standard errors over windows. The seasonal naive forecast uses the value one day before the target.

All ten general-corpus models beat persistence at h=16, 64 and 128 (Table 16). At h=16, skills range from $+0.67$ for Moirai-1.1 to $+0.73$ for Chronos-2, exceeding the seasonal naive forecast's $+0.56$. At h=1, their skills range from $+0.33$ to $+0.40$, except Moirai-1.1 (-0.06). Null underper-

Table 16: ETTh1 skill and the effect of mirror correction.

	Raw forecast skill				Mirror correction, h = 16	
Forecast	h = 1	h = 16	h = 64	h = 128	**Skill**	ρ_{even}
Persistence	0	0	0	0	NE	NA
Seasonal naive (24 h)	−3.842 ± 0.583	+0.563 ± 0.037	+0.467 ± 0.039	+0.490 ± 0.031	NE	NE
Chronos-T5	+0.347 ± 0.038	+0.694 ± 0.023	+0.563 ± 0.032	+0.547 ± 0.023	+0.717	0.02
Chronos-Bolt	+0.358 ± 0.034	+0.697 ± 0.026	+0.573 ± 0.033	+0.567 ± 0.026	+0.712	0.01
Chronos-2	+0.370 ± 0.035	**+0.728** ± 0.022	**+0.584** ± 0.030	+0.577 ± 0.024	+0.724	0.00
TiRex	+0.391 ± 0.031	+0.711 ± 0.025	+0.579 ± 0.032	+0.586 ± 0.027	+0.721	0.01
Moirai-1.1	−0.059 ± 0.059	+0.674 ± 0.022	+0.555 ± 0.027	+0.476 ± 0.040	+0.695	0.04
Moirai-2.0	+0.382 ± 0.032	+0.726 ± 0.023	+0.580 ± 0.031	+0.594 ± 0.025	**+0.725**	0.01
TimesFM-2.0	+0.334 ± 0.039	+0.698 ± 0.023	+0.566 ± 0.030	+0.584 ± 0.024	+0.706	0.01
TimesFM-2.5	+0.402 ± 0.030	+0.718 ± 0.021	+0.581 ± 0.029	+0.583 ± 0.022	+0.718	0.00
Time-MoE	+0.389 ± 0.037	+0.693 ± 0.025	+0.555 ± 0.033	+0.584 ± 0.028	+0.704	0.01
Sundial	**+0.403** ± 0.033	+0.720 ± 0.022	+0.580 ± 0.031	**+0.605** ± 0.025	+0.721	0.01
FinCast	−0.186 ± 0.037	−0.025 ± 0.022	−0.246 ± 0.052	−0.301 ± 0.059	+0.054	0.40

924 windows, context 512, future 128, stride 128. Skill pools squared errors after normalisation by each context's one-step variance. ±: bootstrap SE over windows. Seasonal naive repeats the last observed 24-hour cycle. NE: correction not evaluated for that baseline; NA: even-energy ratio undefined for persistence, whose departure is zero. ρ_{even} retains finite-sample forecast variation. Bold marks the highest skill in each column. This protocol differs from Table 4.

formance thus coexists with useful forecasts on a structured dataset. ETTh1 is a public benchmark; it is not in the three corpora audited in Appendix E.4, but other pretraining corpora may contain it.

FinCast loses to persistence at all four horizons (-0.19, -0.03, -0.25 and -0.30). At h=16, its even-energy ratio is 0.40 and correction raises skill to $+0.05$. For the other models, ratios are at most 0.04 and the corrected skills differ from the raw values by at most 0.02. Most of their departure energy is thus odd on these windows; since the conditions of Proposition 2 are not established for ETTh1, this locates the energy without saying which part is predictive.

F.4 Small-Cap Forecasts and Fitted Drift

The smallest size decile combines positive drift with return autocorrelation, so it tests forecasting on a financial series that has structure.

The equal-weighted smallest size decile of the Kenneth R. French library (downloaded 2026-09-19) has mean context lag-one return autocorrelation $+0.32$ and persistent positive drift; non-synchronous trading is one established explanation for such autocorrelation (Lo & MacKinlay, 1988). The value-weighted largest decile, with autocorrelation $+0.03$, is a comparison.

Each series supplies 120 windows from 1963-07 to 2026-07, with 512 context days, 128 future days, stride 128 and K=4 sign-randomised copies. Unlike the daily anchor, the pooled skill divides each squared return error by the context return variance, reducing the weight of high-volatility windows. Three benchmarks use the context alone: mean return extrapolation, AR(1), and AR(5), with the autoregressions iterated over the horizon. Table 17 gives smallest-decile skill with each forecast's paired difference from the fitted drift, AR(1) ratios, even-energy ratios and the largest-decile comparison at h=16. The AR(1) ratio divides a forecast's raw skill by the fitted AR(1)'s on the smallest decile, so it is not a share of all predictable risk. On the sign-randomised copies the models pool four copies and the fitted AR(1) uses the first, so that column is not ranked.

The drift and autoregressive baselines in Table 17 show that fitted drift accounts for much of the gain over persistence. AR(1)'s point estimate is only 0.021 ± 0.016 higher in the paired comparison. AR(1) falls to -0.031 ± 0.048 on the sign-randomised copies and achieves $+0.067 \pm 0.056$ on the largest decile.

Paired on the same 120 windows, with a cluster bootstrap over future-end dates, the differences from the fitted drift settle the comparison (Table 17). **No model is ahead of the drift by more than two standard errors**; the largest differences are $+0.084 \pm 0.057$ for Moirai-2.0, $+0.079 \pm 0.108$ for Chronos-2 and $+0.048 \pm 0.072$ for TiRex. Time-MoE, Sundial and FinCast are behind it by more than two standard errors. AR(5) is ahead of the drift by $+0.133 \pm 0.059$, and Moirai-1.1, TimesFM-2.0, Time-MoE, Sundial and FinCast are behind AR(5) by more than two standard errors.

Table 17: Forecast skill on the smallest and largest size deciles at $h = 16$.

	Smallest decile (equal-weighted)						Largest decile (value-weighted)
	Raw			Mirror		Sign-randomised	Raw
Forecast	**Skill**	**− drift**	**AR(1) ratio**	**Skill**	ρ_{even}	**Skill**	**Skill**
Persistence	0	NE	0	NE	NA	0	0
Drift (fit)	+0.306 ± 0.059	0	+0.94	NE	NE	NE	**+0.096** ± 0.050
AR(1) (fit)	+0.327 ± 0.060	+0.021 ± 0.016	1	NE	NE	−0.031 ± 0.048	+0.067 ± 0.056
AR(5) (fit)	**+0.439** ± 0.077	+0.133 ± 0.059	**+1.34**	NE	NE	NE	+0.052 ± 0.065
Chronos-T5	+0.299 ± 0.103	−0.007 ± 0.105	+0.91	+0.059	0.40	−0.087	+0.048
Chronos-Bolt	+0.264 ± 0.125	−0.042 ± 0.124	+0.81	+0.328	0.12	−0.116	−0.033
Chronos-2	+0.385 ± 0.112	+0.079 ± 0.108	+1.18	+0.405	0.05	−0.071	+0.009
TiRex	+0.354 ± 0.076	+0.048 ± 0.072	+1.08	+0.368	0.11	−0.084	−0.074
Moirai-1.1	+0.259 ± 0.083	−0.047 ± 0.062	+0.79	−0.024	1.07	−0.191	−0.189
Moirai-2.0	+0.391 ± 0.070	+0.084 ± 0.057	+1.19	**+0.427**	0.09	−0.101	−0.066
TimesFM-2.0	+0.205 ± 0.083	−0.101 ± 0.087	+0.63	+0.195	0.23	−0.189	−4.643
TimesFM-2.5	+0.277 ± 0.135	−0.029 ± 0.115	+0.85	+0.360	0.07	−0.066	−0.036
Time-MoE	−5.172 ± 2.445	−5.478 ± 2.481	−15.81	−1.691	0.31	−0.991	−1.641
Sundial	−0.399 ± 0.140	−0.705 ± 0.158	−1.22	−0.273	0.21	−0.255	−0.558
FinCast	+0.036 ± 0.096	−0.270 ± 0.094	+0.11	−0.107	0.70	−0.057	+0.012

120 windows per series since 1963; context 512, requested future 128; skill pools squared return errors divided by each context's return variance. − drift: paired skill difference against the fitted drift (not bolded). ±: future-end-date cluster-bootstrap SE where shown. NE: not evaluated; NA: undefined for persistence. Bold marks the highest raw or mirror skill, or AR(1) ratio, within each column, not across portfolio constructions.

Nine models have positive raw-window skill, from FinCast's $+0.036 \pm 0.096$ to Moirai-2.0's $+0.391 \pm 0.070$. Chronos-2, TiRex and Moirai-2.0 exceed AR(1)'s point estimate, but the paired comparisons put none of them more than two standard errors above fitted drift. Time-MoE and Sundial have negative skill (-5.172 and -0.399). All eleven models lose to persistence on the copies, and only three have positive point estimates on the largest decile. At h=1, eight models' departures correlate positively with the fitted AR(1)'s departure by more than one standard error, but none achieves positive skill.

Removing the even part raises skill for seven models and lowers it for Chronos-T5, Moirai-1.1, TimesFM-2.0 and FinCast. After mirror correction, the h=16 skills of Chronos-Bolt, Chronos-2, TiRex, Moirai-2.0 and TimesFM-2.5 are $+0.328$, $+0.405$, $+0.368$, $+0.427$ and $+0.360$, at or above their raw point estimates. Their even-energy ratios are 0.05–0.12. Moirai-1.1 falls from $+0.259$ to -0.024, with ratio 1.07, and Chronos-T5 falls from $+0.299$ to $+0.059$, with ratio 0.40. The ratio exceeds one because odd and even parts need not be orthogonal on raw data; Proposition 2(iii) accounts for these changes.

F.5 Falling Markets and Remembered History

Design. The daily anchor's windows end between 2017 and 2026, when equities rose. To test the interpretation of Section 5 where markets fell, we built a second anchor and wrote down four predictions before running any model on it. The windows follow the construction of Appendix F.1 on the same 50 equity series, the market and the 49 industry portfolios, from 1994 to 2014: 512-day contexts and 128-day futures ending on 37 dates from 1996-09-10 to 2014-12-31, with one sign-randomised copy per window, 1850 windows in all. A window is *down* when the market fell over its future, which happens for 8 end dates and 400 windows (futures ending in 1998-09, 2000-09, 2001-04, 2002-10, 2007-11, 2008-05, 2008-11 and 2011-12), and *up* otherwise. From their N1 results we classed Chronos-T5, Chronos-Bolt, Chronos-2, TiRex and Moirai-1.1 as upward-biased and predicted, for each of them, that raw skill at h=128 is lower on down windows than on up windows (P1) and that the mirror correction lowers skill on up windows and raises it on down windows (P2); that the mean up-minus-down gap in raw skill is larger for this group than for Moirai-2.0, TimesFM-2.0, TimesFM-2.5, Time-MoE and Sundial, with FinCast reported separately (P3); and that every model loses to the last value on the sign-randomised copies in both regimes (P4). A prediction counts as met when the point estimates go the stated way. Every model forecasts each raw window, its copy and its multiplicative mirror, as in Appendix F.1.

Outcome. Table 18 gives the results. P3 and P4 hold. The mean up-minus-down gap is $+0.24$ for the upward-biased group and -0.33 for the five others (-0.27 with FinCast); the second mean comes from Time-MoE (-1.04) and Sundial (-0.64), and without them it is $+0.01$. Per model, all five upward-biased gaps are positive ($+0.07$ to $+0.45$, median $+0.20$), against one of the five others (median -0.11), and in 23 of the 25 cross-group pairs the upward-biased model has the larger gap (exact two-sided Mann–Whitney $p = 0.03$, a test not fixed in advance). Every model loses on the copies. P1 holds for all five upward-biased models. P2 holds for Chronos-T5 and Moirai-1.1: the correction changes their skill by -0.131 ± 0.026 and -0.110 ± 0.034 on up windows and by $+0.064 \pm 0.050$ and $+0.208 \pm 0.062$ on down windows, and the alignment term of Proposition 2(iii) changes sign with the market (-0.19 and $+0.05$ for Chronos-T5, -0.38 and $+0.13$ for Moirai-1.1). Their even parts are upward drifts in both regimes, $+4.8$ and $+3.0$ percent a year for Chronos-T5 and $+9.0$ and $+7.4$ for Moirai-1.1, which is the interpretation of Section 5; Chronos-T5 does not time the market and Moirai-1.1 does so only weakly (0.21 and 0.34; Table 19). P2 fails for Chronos-Bolt, Chronos-2 and TiRex. For Chronos-2 the premise is missing: on these windows its even part is a downward drift (-1.2 and -3.8 percent a year), so by part (iii) the correction helps where prices rose and hurts where they fell, as observed. For Chronos-Bolt and TiRex the even part is lower before falls ($+7.6$ against -3.5 and $+5.8$ against $+1.5$ percent a year), for the reason given next. The calendar split by the two bear markets gives the same picture for Moirai-1.1, Chronos-Bolt and TiRex.

Remembered history. Several models forecast these older windows better than simple readings of their contexts can. We averaged each model's h=128 forecast return over the windows that end on the same date and correlated these means with the mean realised return of the same windows across the 37 dates (Figure 7). Chronos-Bolt, TiRex, Moirai-2.0 and TimesFM-2.0 reach 0.71 to 0.87 (Table 19), with slopes of 0.29 to 0.62, and still 0.58 to 0.78 without the crash of 2008, while rules that read only the context, the fitted drift or momentum over 20 to 511 days, followed or reversed, reach at most 0.26; all four beat the last value during the two bear markets ($+0.22$ to $+0.39$). Sundial, Time-MoE, Moirai-1.1 and FinCast also exceed the best rule, but only at 0.41, 0.35, 0.34 and 0.29. From contexts that end on 2008-05-21, Chronos-Bolt forecasts falls of 74 and 61 percent over the next 128 days for the mining and steel portfolios, which fell 73 and 75 percent, and TimesFM-2.0 falls of 79 and 62 percent. The timing does not carry over to the other two passes over the same windows. From the mirrored contexts Chronos-Bolt, TiRex and Moirai-2.0 reach -0.17 to -0.02, and from the sign-randomised copies, which keep each context's magnitudes, 0.25 to 0.31. A forecast that responds to the direction of the context would reverse on the mirror, and one that responds to its magnitudes would survive the randomisation; **a forecast that recalls the series would do neither, which is what these three models show**. A mixture of the first two would need a direction-reading part that times the market better than any of the trend rules. TimesFM-2.0's timing instead reverses on the mirror (-0.82), so whatever it recalls is reversed with the history, and its forecast paths follow the realised day-to-day moves more often than chance allows (median correlation of increments $+0.09$, against 0.00 on the copies).

A second check asks whether the forecasts follow the calendar. We moved the end of every context by 20 or 40 trading days in either direction, 7300 further windows per model. Such a shift changes 4 to 8 percent of a 512-day context but moves the 128-day target date by as much. Within each end date we correlate, across the five shifts, the mean forecast change from each shifted context's end with the mean realised level at its target date, both relative to the unshifted context's end, and pool the deviations from each date's mean over dates; the p-value multiplies each date's realised deviations by a random sign. The realised deviations then depend only on moves at least 48 days after the latest shifted context end. On simulated random walks, forecasts that read only their contexts average between -0.01 and $+0.01$ whether they follow or oppose recent moves, and a forecast that recalls the path averages 0.64 to 0.71 (Figure 11). Correlating returns instead would be biased, because a later shift's context contains part of the earlier shifts' futures: a forecast that opposes the last 20 days' move then scores $+0.35$ and one that follows it -0.35. When the drift itself changes sign, forecasts that read the context can score too (up to 0.46 in absolute value in the simulation), so Table 20 also gives the fitted drift and momentum over 20 to 250 days, followed or reversed. On 1996–2014, TimesFM-2.0 scores $+0.58$ ($p < 10^{-3}$), Chronos-Bolt $+0.41$, Moirai-2.0 $+0.33$ and TiRex $+0.29$ (p from 0.04 to 0.11), Moirai-1.1 and Chronos-2 -0.05 and $+0.11$, and the best context rule 0.29 in absolute value. The check is thus clear for TimesFM-2.0, points the same way for Chronos-Bolt, and cannot separate Moirai-2.0 and TiRex from a rule that opposes the context's long-run trend. With

the timing check, it indicates that these models, TimesFM-2.0 most clearly, saw these years during pretraining, in these series or in series closely tied to them; only their training data could confirm it.

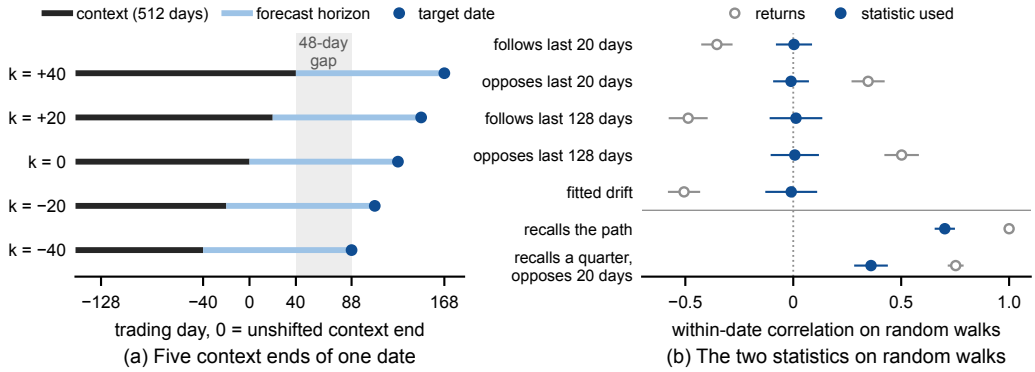


Figure 11: **The calendar-shift check pairs each forecast change with the realised level at its target date.** (a) For one end date, the five context ends (k from -40 to $+40$ trading days) and their 128-day horizons; every target date lies at least 48 days after the latest context end, so the target deviations within a date depend only on moves that no context contains. (b) Within-date correlation on simulated random walks (37 dates of 50 series; bars ± 1 standard deviation over 200 replications, 100 for the two recalling forecasts), with returns from each shifted context end (open) and with the statistic used (filled). Returns score forecasts that only react to the context at up to ± 0.5; the statistic used scores them near zero and still detects recall.

We ran the same check on the paper's own anchors after predicting that none of the four models would follow the calendar on the daily anchor, with no prediction for the size deciles. All shifts, 6260 daily-anchor and 1192 size-decile windows per model, reproduce the published forecasts at zero shift (up to sampling for Moirai-1.1). No model follows the calendar on either anchor: the largest correlations are $+0.27$ on the daily anchor's equities, $+0.09$ on its exchange rates and $+0.18$ on the deciles, while momentum over 60 days reaches 0.50 on the anchor's equities, so the statistic can register structure there (Table 20). The predictions were written for the return version of the statistic; by that version TiRex and TimesFM-2.0 would be flagged on exchange rates ($+0.53$ and $+0.54$), which the unbiased version does not reproduce (-0.04 and $+0.09$).

The remembered series in public training data. We then searched the public pretraining data of the corpus audit (Appendix E.4) for the market itself, after writing down predictions. For each of the 287,251 series of the 213 non-synthetic subsets, we took the largest correlation between its daily log returns and those of the Ken French market over all alignments with an overlap of at least 250 days; chance maxima are near 0.3 at this overlap. 79 series exceed 0.9 (0.90 to 0.98), **all in the copies of M4 daily in the three corpora** and 69 of them distinct; each stays above 0.85 without the five largest market days. Their spans run from 1984 to October 2017 and cover 82 percent of the trading days of 1996–2014, the crash of 2008 included, and none after October 2017. M4 daily is part of the training data of Chronos, whose in-domain datasets contributed their training portions, and of TimesFM, and it is in LOTSA and Time-300B; TiRex's model card lists the Chronos datasets, Moirai-2.0's lists the training portions of the GIFT-Eval datasets, which include M4, and Chronos-Bolt's lists none. Chronos-T5, Moirai-1.1 and Time-MoE were trained on these corpora; Chronos-T5 does not time the market, and the other two do so far less than the four models (0.34 and 0.35 against 0.71 to 0.87; Table 19), so holding the series in the training data does not by itself make a model reproduce it. We had predicted that the covered days would include none of the daily anchor's futures; they include 77 percent of the future of its earliest equity end date and nothing of the other 17. We had also planned to compare the timing on covered and uncovered end dates, but only 6 of the 37 are uncovered, fewer than the 8 we required, so we report the coverage alone.

On a window a model remembers, the raw forecast carries the remembered move and the mirrored forecast does not, so the even part is half of that move and is aligned with it whichever way the market went. Removing it then lowers skill in both regimes, which is how P2 failed for Chronos-Bolt and TiRex: the correction cannot separate a bias from a memory. On the 2017–2026 anchor

the raw-context timing is at most 0.40, with slopes of at most 0.19, below what 60-day momentum reaches there (0.50) and what the mirrored and sign-randomised passes reach with only 18 dates (up to 0.65), and no model follows the calendar. That anchor's interpretation therefore stands, but it rests most safely on the windows that end after 2025, past the training cut-offs of the 2024 models, where Chronos-T5's skill is $+0.074 \pm 0.045$ raw and $+0.011 \pm 0.043$ corrected (Table 24); Chronos-Bolt's and TiRex's raw equity skills of about $+0.12$ fall to near zero there.

Table 18: Falling-market test at h=128.

	Raw skill		Mirror change		Even part (%/yr)		Null
Model	**Up**	**Down**	**Up**	**Down**	**Up**	**Down**	**skill**
Chronos-T5	+0.219 ± 0.048	−0.151 ± 0.085	−0.131 ± 0.026	+0.064 ± 0.050	+4.8	+3.0	**−0.076**
Chronos-Bolt	+0.326 ± 0.099	+0.254 ± 0.315	−0.119 ± 0.047	−0.016 ± 0.160	+7.6	−3.5	−0.113
Chronos-2	−0.040 ± 0.131	−0.166 ± 0.064	+0.055 ± 0.075	−0.156 ± 0.044	−1.2	−3.8	−0.097
TiRex	+0.312 ± 0.072	+0.111 ± 0.170	−0.130 ± 0.029	−0.050 ± 0.060	+5.8	+1.5	−0.145
Moirai-1.1	+0.171 ± 0.057	−0.277 ± 0.096	−0.110 ± 0.034	+0.208 ± 0.062	+9.0	+7.4	−0.150
Moirai-2.0	+0.123 ± 0.140	+0.229 ± 0.215	+0.078 ± 0.066	−0.063 ± 0.117	+3.8	−7.9	−0.160
TimesFM-2.0	**+0.358** ± 0.097	**+0.384** ± 0.227	−0.025 ± 0.020	−0.011 ± 0.023	+1.6	+0.1	−0.198
TimesFM-2.5	+0.023 ± 0.048	−0.153 ± 0.033	+0.066 ± 0.045	−0.008 ± 0.031	+0.3	−0.2	−0.090
Time-MoE	−1.183 ± 0.235	−0.147 ± 0.145	+0.015 ± 0.189	+0.021 ± 0.066	+8.3	+4.7	−1.038
Sundial	−0.581 ± 0.144	+0.055 ± 0.076	−0.078 ± 0.129	−0.018 ± 0.064	+7.7	+4.3	−0.569
FinCast	+0.022 ± 0.047	−0.016 ± 0.018	+0.008 ± 0.022	+0.015 ± 0.021	+0.4	+1.2	−0.084

1850 equity windows with futures ending 1996–2014, context 512, future 128. Up and down: the market's return over the window's future is non-negative (1450 windows) or negative (400). Raw skill pools squared return errors; mirror change is mirror-corrected minus raw skill; ±: cluster-bootstrap SE over the 37 end dates. Even part: $100(252/128)\mathbb{E}[a]$ at h=128. Null skill: the sign-randomised copies, all windows. Rows: the models classed as upward-biased before the run, the others, and FinCast, reported separately. Bold marks the highest raw skill in each regime and the highest null skill.

Table 19: Market timing from raw, mirrored and sign-randomised contexts at h=128.

	1996–2014, 37 end dates			2017–2026 equities, 18 end dates		
Model	**Raw**	**Mirror**	**Randomised**	**Raw**	**Mirror**	**Randomised**
Chronos-T5	+0.21	+0.17	+0.04	−0.22	+0.25	−0.28
Chronos-Bolt	**+0.78***	−0.02	+0.29	+0.40	+0.16	+0.28
Chronos-2	−0.15	+0.31	+0.19	−0.13	+0.03	+0.53
TiRex	**+0.71***	−0.17	+0.31	+0.39	+0.50	+0.37
Moirai-1.1	**+0.34***	+0.09	+0.20	−0.00	+0.62	+0.65
Moirai-2.0	**+0.76***	−0.05	+0.25	+0.16	+0.18	+0.24
TimesFM-2.0	**+0.87***	−0.82	−0.09	+0.31	+0.09	+0.42
TimesFM-2.5	+0.02	−0.03	+0.01	+0.04	−0.01	+0.35
Time-MoE	**+0.35***	−0.27	+0.14	+0.23	+0.20	+0.39
Sundial	**+0.41***	−0.32	+0.10	+0.12	+0.37	+0.43
FinCast	+0.29	−0.36	+0.06	+0.34	+0.06	−0.15
Context rule, largest $\lvert r \rvert$	0.26	–	–	0.50	–	–

Correlation across end dates between the mean forecast return and the mean realised return of the raw windows ending on that date. Raw: forecasts from the contexts; mirror: from their multiplicative mirrors; randomised: from their sign-randomised copies, all scored against the raw windows' realised returns. *: permutation $p < 0.01$ over dates, raw forecasts only. Context rule: the largest absolute correlation of the fitted drift or of momentum over 20, 60, 128, 250 or 511 days, followed or reversed. The 2017–2026 columns use the 900 equity windows of the daily anchor. Bold marks a starred correlation above the context rule of its column.

F.6 Directional Accuracy

Studies of financial forecasts often report directional accuracy, the share of windows in which the forecast moves from the last value the way the price then moved. On a rising market it rewards the directional bias of Section 3. We computed it from the published forecasts after writing down predictions. On the daily anchor's raw equity windows 69 percent rose over 128 days, so **always forecasting a rise scores 69 ± 6 percent, more than every model** (40 to 66 percent; Table 21), and across the eleven models the score tracks how often the model forecasts a rise (correlation 0.96). The mirror correction lowers the score of all five models whose N1 forecasts lean upward, TiRex's from 66 to 58 percent. On the sign-randomised copies rises are slightly rarer than falls, because compounding lowers the median (47 percent at h=128), and the models score 48 to 50 percent. We had predicted that every model would be within two standard errors of one half there; the prediction fails for Chronos-T5, TiRex and Moirai-1.1, mostly for this reason, although Chronos-

Table 20: Calendar-shift check at h=128: forecast change against realised target level.

	1996–2014	Daily anchor, 2017–2026		Size deciles, 1963–2026	
Forecast	**Equities**	**Equities**	**Exchange rates**	**Smallest**	**Largest**
Chronos-Bolt	+0.41	+0.27	−0.09	+0.18	+0.12
TiRex	+0.29	−0.14	−0.04	−0.01	+0.06
Moirai-2.0	+0.33	+0.12	0.00	+0.01	+0.16
TimesFM-2.0	**+0.58***	−0.17	+0.09	+0.08	+0.12
Moirai-1.1	−0.05	+0.05	−0.11	+0.07	+0.08
Chronos-2	+0.11	−0.03	−0.17	−0.04	+0.04
Context rule, largest $\lvert r\rvert$	0.29	0.50	0.16	0.12	0.06

Each context end is moved by $k \in \{-40, -20, 0, +20, +40\}$ trading days, keeping 512 context and 128 future days. Per end date and shift: the mean forecast change from the shifted context's end, and the mean realised level at the target date relative to the unshifted context's end, over 50 equity series, 20 exchange rates or one decile series. Entries correlate their deviations from each date's mean over its shifts, pooled over 37, 18, 19, 120 and 120 end dates. *: $p < 0.01$, random sign per date on the realised deviations. Context rule: the largest absolute correlation of the fitted drift or of momentum over 20, 60, 128 or 250 days, followed or reversed, without a significance claim. Bold marks a starred correlation above the context rule of its column.

Table 21: Directional accuracy at h=128 follows how often a model forecasts a rise.

	Daily anchor, equities, 2017–2026			Randomised	1996–2014
Forecast	**Raw**	**Upward**	**Mirror-corrected**	**All copies**	**Raw**
Always up	**69** ± 6	100	–	47 ± 1	69 ± 4
Sign of fitted drift	63 ± 5	–	–	–	59 ± 4
Chronos-T5	61 ± 5	84	57 ± 3	49 ± 1	65 ± 4
Chronos-Bolt	66 ± 5	88	61 ± 4	**50** ± 1	**72** ± 3
Chronos-2	64 ± 5	85	**62** ± 4	**50** ± 1	64 ± 4
TiRex	66 ± 4	78	58 ± 3	48 ± 1	69 ± 3
Moirai-1.1	59 ± 3	70	50 ± 3	48 ± 1	63 ± 3
Moirai-2.0	62 ± 4	74	61 ± 4	**50** ± 1	69 ± 3
TimesFM-2.0	58 ± 3	60	54 ± 3	49 ± 1	**72** ± 3
TimesFM-2.5	56 ± 3	62	55 ± 4	49 ± 1	57 ± 2
Time-MoE	41 ± 5	26	38 ± 5	49 ± 1	44 ± 4
Sundial	40 ± 5	23	36 ± 5	49 ± 1	44 ± 4
FinCast	56 ± 4	47	48 ± 3	**50** ± 1	49 ± 3

Percentage of windows in which the forecast and the realised 128-day move from the last value have the same sign (ties count one half); ±: cluster-bootstrap SE over family and end date. Daily anchor: the 900 raw equity windows; Upward: percentage of forecasts above the last value; Mirror-corrected: the sign of the odd part. All copies: the four sign-randomised copies of all 1280 windows, each scored against its own future. 1996–2014: the 1850 falling-market windows. Bold marks the highest accuracy in each accuracy column, baselines included.

T5's 52 percent at h=16 is not explained by it. On the 1996–2014 windows TimesFM-2.0 and Chronos-Bolt, which time the market, score 72 percent against 69 for always forecasting a rise.

G Symmetry Correction and Its Failure Cases

G.1 Transform Sampling and Evaluation Design

Scoring independent futures. Every model forecasts, at $H{=}128$, the same 256 contexts per rung: 128 draws of Section 2.2's generators and their mirrors about y_0, future included, so that the mirror correction needs no extra pass (row i's mirror is row $i + 128$). Only the 128 draws are scored, each against its own future; the mirror rows supply $D(\bar{x})$ and enter nothing else.

Scoring both members of every mirrored pair would impose a risk improvement algebraically: on a test set that holds both members of a pair, the raw and mirrored errors are $(E + O - \Delta)^2$ and $(E - O + \Delta)^2$ with E and O the even and odd parts of the departure and Δ the future increment, their mean is $E^2 + (O - \Delta)^2$, and the corrected error is $(O - \Delta)^2$, so the mirror's risk reduction on such a set equals $\mathbb{E}[E^2]$ whatever the data. On the scored half the reduction is $\mathbb{E}[E^2] + 2\,\mathbb{E}[E(O - \Delta)]$, whose second term is zero in expectation under the symmetry and is estimated from the original futures.

Transform draws and uncertainty. The sign-averaged correction uses $K{=}4$ independent sign draws per context, 1024 extra forecasts per rung and model. On the anchor the mirror of every raw window is forecast once ($\bar{x}_i = \bar{x}_{i-1}(1 - r_i)$) and the $K{=}4$ surrogates of Appendix F.1 serve as the sign draws, and the corrected raw-window forecasts are in return units, $\hat{y}/y_T - 1 - a$, with the same pooled skill as the daily evaluation.

Standard errors are a bootstrap over series on the ladder and the anchor's cluster bootstrap on the windows. The even-energy ratio is $\mathbb{E}[a^2]/\mathbb{E}[D^2]$ for the mirror, from the same forecasts.

The sign-average ratio subtracts the estimated across-draw variance divided by K from its numerator; the reported $K \to \infty$ skill subtracts that variance term from the corrected squared error. Both are four-draw estimates, and sampled model outputs contribute to the across-draw variation.

G.2 Mirror Algorithm and Context-Fitted Baselines

Algorithm 1 uses two forecasts, one on the context and one on its mirror, to form $D \mapsto \frac{1}{2}(D(x) - D(\bar{x}))$, the two-element group average of Chen et al. (2020). It needs only a forecast interface that accepts transformed contexts; Appendix G.8 compares it with mirror augmentation during fine-tuning.

The optional gate in Algorithm 1 uses a constant-departure benchmark to decide whether to keep the raw forecast. It ignores the variation and covariance terms of Equation 5, and $h\hat{\mu}$ only approximates the multistep drift.

Algorithm 1 Mirror-averaged correction of a frozen forecaster (price levels, return units)

Require: context $x_0, \ldots, x_T$; frozen forecaster g; horizon h; optional estimated drift $\hat{\mu}$ per step

$r_i \leftarrow x_i/x_{i-1} - 1$ for $i = 1, \ldots, T$ {returns of the context}

$\bar{x}_0 \leftarrow x_0$; $\bar{x}_i \leftarrow \bar{x}_{i-1}(1 - r_i)$ {the mirror: every return negated}

$\hat{y} \leftarrow g(x)_h$; $\hat{y}^{\,m} \leftarrow g(\bar{x})_h$ {$g(\cdot)_h$: point forecast h steps ahead; the second pass is the extra one}

$D \leftarrow \hat{y}/x_T - 1$; $D^m \leftarrow \hat{y}^{\,m}/\bar{x}_T - 1$ {departures in return units}

$a \leftarrow \frac{1}{2}(D + D^m)$ {the even, sign-invariant part}

if $\hat{\mu}$ is given **and** $a\,(2\hat{\mu}h - a) > 0$ **then**

  return $\hat{y}$ {optional constant-benchmark heuristic}

end if

return $\hat{y}^{\mathrm{c}} \leftarrow x_T\,(1 + D - a)$ {only the odd part of the forecast survives}

For additive series replace returns by increments, $\bar{x}_i = \bar{x}_{i-1} - (x_i - x_{i-1})$, and departures by $\hat{y} - x_T$; for the sign-averaged variant draw K sign vectors instead of the one mirror and average the K departures, at the finite-K cost of Proposition 4(a). Except for Table 27, the reported correction is unconditional. That table evaluates the optional gate with $\hat{\mu}$ set to the context's mean return.

Context-fitted baselines. The final rows of Table 1 use only each context: its mean increment or return, and an AR(1) fitted by least squares and iterated over the horizon. Both have negative null skill. On the additive ladder their departures are odd, so mirroring leaves them unchanged; they capture none and 53 percent of the N4 optimum's gain, respectively. On multiplicative financial windows, compounding can introduce an even component.

G.3 Correction on the Synthetic Ladder

On N1 at h=128, sign averaging changes Chronos-T5's mean departure from $+3.24 \pm 0.32\sigma$ to $+0.46 \pm 0.32$, Moirai's from $+2.34 \pm 0.41$ to $+0.09 \pm 0.44$, and FinCast's from -0.37 ± 0.20 to $+0.14 \pm 0.21$. N2 and N3 show similar reductions (Table 22). Mirror averaging lowers estimated risk in 135 of 176 model–rung–horizon cells, raises it in 25, and leaves TimesFM-2.5's 16 cells unchanged. Of the gains, 77 exceed one standard error of the paired difference; 7 losses do so. These are descriptive counts over dependent comparisons. Chronos-T5's N1 skill at h=16 rises by 0.071 ± 0.023. The larger losses occur for TimesFM-2.0, Chronos-2 and Sundial at long N4 horizons, Time-MoE at N4's first step, and Chronos-2 and Sundial in one N3 cell each.

At h=16 on N1 the even-energy ratios are 0.48 for Chronos-T5, 0.79 for Moirai-1.1, 0.55 for FinCast, 0.15 for TimesFM-2.0 and 0.04 for Time-MoE.

With K=4, the sign average must offset its sampling cost. Chronos-T5's corrected ratio of 0.33 on N1 at h=128 exceeds the theoretical 0.20 threshold for a fixed predictor, and its skill improves from -0.229 to -0.140. Chronos-Bolt and TiRex also improve in this setting. On N4, independent sign randomisation destroys dependence and has no corresponding guarantee; it lowers every model's fraction of the optimum's gain, to -0.30 for Moirai.

TimesFM-2.5 already averages over additive negation, so the extra mirror leaves its ladder forecasts unchanged, and Moirai-2.0's N1 skill declines within one standard error despite an even-energy ratio of 0.54.

G.4 Correction on Randomised and Raw Daily Returns

Mirror correction improves h=128 skill for all eleven models on the sign-randomised daily data (Table 23). Chronos-T5, Moirai, TimesFM-2.0 and FinCast have even-energy ratios of 0.40–0.56. At K=4, sign averaging loses skill for ten of the eleven models. Subtracting the estimated four-draw sampling variance (the $K \to \infty$ estimate) puts skill above the raw forecast for nine models and at most 0.002 below it for Chronos-2 and Moirai-2.0. TimesFM-2.5 retains an even-energy ratio of 0.25 on the surrogates: its internal additive averaging differs from the multiplicative mirror used here.

On raw windows the same correction can remove useful forecasts. Chronos-T5's mean departure falls from $+2.17$ to $+0.64\sigma_r$ and its pooled skill from $+0.04 \pm 0.06$ to -0.02 ± 0.04, with an even-energy ratio of 0.49. Its post-2025 point estimate also falls, from $+0.07$ to $+0.01$. Every other model except Moirai-2.0 also loses raw-window skill, among them Chronos-Bolt and TiRex (even-energy ratios 0.34 and 0.62). Moirai-2.0 improves from -0.01 ± 0.09 to $+0.03 \pm 0.05$; Sundial falls from -0.54 ± 0.20 to -0.95 ± 0.30; Appendix G.6 accounts for these changes. At h=16 the correction instead raises raw-window skill on equities for eight models and on exchange rates for seven.

G.5 Sampling Noise in the Mirror's Gain

For an output that is the mean of S sampled paths, write $\hat{y}_S(x) = \hat{y}_\infty(x) + \epsilon(x)$ with $\mathbb{E}[\epsilon \mid x] = 0$ and $\mathrm{Var}(\epsilon \mid x) = v(x)/S$, drawn independently for x and its mirror $\bar{x}$. The raw forecast then carries $\mathbb{E}[v(x)]/S$ of sampling variance in its risk and the mirror-corrected forecast $\frac{1}{4}\mathbb{E}[v(x) + v(\bar{x})]/S$, so part of the mirror's gain is the averaging of two independent sample means rather than the removal of an even component of $\hat{y}_\infty$. Table 25 computes that part from the within-run estimates of v/S stored with every forecast and sets it against the observed change in skill, on the scored half of the ladder and on the daily windows in return units. For Chronos-T5 and Sundial it is at most a tenth of every gain. For Moirai-1.1 it is all of the small N1 gain at h=16, a tenth to a third of the gains at h=128 and 13 percent of the gain on the randomised daily windows. It is negligible on the raw

Table 22: Synthetic-ladder skill before and after symmetry correction.

		Skill, h = 16			Skill, h = 128		
Model	**Rung**	**Raw**	**Mirror**	**Sign avg.**	**Raw**	**Mirror**	**Sign avg.**
Chronos-T5	N1	−0.139 ± 0.044	**−0.068** ± 0.031	−0.112 ± 0.042	−0.229 ± 0.078	**−0.076** ± 0.042	−0.140 ± 0.056
	N2	−0.093 ± 0.048	**−0.039** ± 0.035	−0.090 ± 0.051	−0.170 ± 0.066	**−0.033** ± 0.044	−0.111 ± 0.061
	N3	−0.023 ± 0.033	**−0.015** ± 0.033	−0.039 ± 0.043	−0.176 ± 0.058	**−0.096** ± 0.049	−0.183 ± 0.063
	N4	−0.196 ± 0.112	**−0.095** ± 0.096	−0.179 ± 0.116	−0.932 ± 0.213	**−0.319** ± 0.107	−0.794 ± 0.159
Chronos-Bolt	N1	−0.098 ± 0.059	−0.060 ± 0.040	**−0.058** ± 0.053	−0.224 ± 0.092	**−0.050** ± 0.039	−0.176 ± 0.064
	N2	−0.057 ± 0.079	**+0.009** ± 0.042	−0.018 ± 0.078	−0.247 ± 0.099	**−0.040** ± 0.041	−0.186 ± 0.098
	N3	−0.072 ± 0.064	−0.054 ± 0.044	**−0.043** ± 0.055	−0.137 ± 0.063	**−0.100** ± 0.040	−0.191 ± 0.060
	N4	−0.122 ± 0.093	**−0.017** ± 0.067	−0.127 ± 0.104	−0.617 ± 0.242	**−0.096** ± 0.089	−0.480 ± 0.179
Chronos-2	N1	−0.021 ± 0.024	**−0.012** ± 0.022	−0.043 ± 0.029	−0.053 ± 0.037	**−0.043** ± 0.034	−0.047 ± 0.037
	N2	**−0.013** ± 0.036	−0.018 ± 0.036	**−0.013** ± 0.039	−0.062 ± 0.058	**−0.016** ± 0.055	−0.062 ± 0.064
	N3	−0.100 ± 0.105	−0.114 ± 0.118	**−0.091** ± 0.120	**−0.022** ± 0.075	−0.100 ± 0.145	−0.090 ± 0.084
	N4	+0.029 ± 0.060	+0.025 ± 0.059	**+0.048** ± 0.068	**−0.018** ± 0.046	−0.048 ± 0.049	−0.142 ± 0.096
TiRex	N1	−0.097 ± 0.043	**−0.045** ± 0.026	−0.062 ± 0.039	−0.098 ± 0.066	**+0.005** ± 0.038	−0.040 ± 0.053
	N2	−0.066 ± 0.050	−0.039 ± 0.030	**−0.028** ± 0.063	−0.170 ± 0.079	**−0.021** ± 0.047	−0.045 ± 0.056
	N3	−0.051 ± 0.041	**−0.023** ± 0.029	−0.056 ± 0.044	−0.097 ± 0.056	**−0.064** ± 0.042	−0.139 ± 0.052
	N4	+0.053 ± 0.065	**+0.057** ± 0.068	−0.062 ± 0.106	−0.116 ± 0.075	**−0.101** ± 0.072	−0.600 ± 0.176
Moirai-1.1	N1	−0.059 ± 0.061	−0.048 ± 0.050	**−0.018** ± 0.071	−0.212 ± 0.105	**−0.103** ± 0.054	−0.195 ± 0.102
	N2	−0.265 ± 0.074	**−0.113** ± 0.052	−0.340 ± 0.090	−0.197 ± 0.098	**−0.097** ± 0.052	−0.154 ± 0.084
	N3	−0.270 ± 0.083	**−0.193** ± 0.064	−0.270 ± 0.092	−0.200 ± 0.085	**−0.059** ± 0.052	−0.137 ± 0.075
	N4	−0.157 ± 0.089	**−0.036** ± 0.069	−0.276 ± 0.138	−0.767 ± 0.284	**−0.224** ± 0.121	−0.987 ± 0.307
Moirai-2.0	N1	**−0.037** ± 0.055	−0.045 ± 0.044	−0.075 ± 0.060	−0.128 ± 0.078	**−0.095** ± 0.070	−0.201 ± 0.093
	N2	**+0.007** ± 0.047	−0.026 ± 0.040	−0.029 ± 0.055	−0.123 ± 0.066	**−0.052** ± 0.059	−0.071 ± 0.069
	N3	−0.166 ± 0.064	**−0.110** ± 0.045	−0.190 ± 0.080	−0.218 ± 0.104	**−0.090** ± 0.063	−0.319 ± 0.115
	N4	−0.010 ± 0.067	**−0.008** ± 0.067	−0.230 ± 0.127	−0.058 ± 0.066	**−0.051** ± 0.061	−0.475 ± 0.164
TimesFM-2.0	N1	−0.186 ± 0.179	**−0.096** ± 0.096	−0.232 ± 0.206	−0.057 ± 0.058	**−0.021** ± 0.051	−0.087 ± 0.073
	N2	−0.087 ± 0.056	**−0.077** ± 0.067	−0.141 ± 0.062	**−0.224** ± 0.099	−0.258 ± 0.093	−0.239 ± 0.090
	N3	−0.139 ± 0.068	**−0.133** ± 0.054	−0.162 ± 0.064	−0.209 ± 0.083	**−0.147** ± 0.062	−0.280 ± 0.088
	N4	−0.014 ± 0.075	**+0.010** ± 0.071	−0.166 ± 0.114	**−0.034** ± 0.070	−0.058 ± 0.063	−0.211 ± 0.108
TimesFM-2.5	N1	**−0.023** ± 0.030	**−0.023** ± 0.030	−0.056 ± 0.034	**−0.011** ± 0.027	**−0.011** ± 0.028	−0.029 ± 0.032
	N2	−0.020 ± 0.053	−0.020 ± 0.054	**−0.011** ± 0.059	**−0.032** ± 0.063	**−0.032** ± 0.064	−0.072 ± 0.065
	N3	−0.057 ± 0.035	−0.057 ± 0.036	**−0.046** ± 0.038	**−0.038** ± 0.041	**−0.038** ± 0.041	−0.072 ± 0.051
	N4	**+0.040** ± 0.065	**+0.040** ± 0.069	+0.010 ± 0.077	**−0.073** ± 0.050	**−0.073** ± 0.051	−0.170 ± 0.072
Time-MoE	N1	−0.554 ± 0.262	**−0.511** ± 0.251	−0.718 ± 0.305	−0.723 ± 0.217	**−0.722** ± 0.220	−0.870 ± 0.240
	N2	−0.495 ± 0.334	**−0.471** ± 0.323	−0.669 ± 0.329	−0.718 ± 0.242	**−0.685** ± 0.230	−0.749 ± 0.234
	N3	−1.717 ± 0.430	**−1.420** ± 0.338	−2.010 ± 0.430	−1.296 ± 0.307	**−1.206** ± 0.268	−1.540 ± 0.299
	N4	−0.243 ± 0.176	**−0.190** ± 0.150	−1.249 ± 0.365	−0.782 ± 0.258	**−0.717** ± 0.232	−2.094 ± 0.464
Sundial	N1	−0.179 ± 0.089	**−0.129** ± 0.088	−0.207 ± 0.108	−0.466 ± 0.158	**−0.444** ± 0.151	−0.569 ± 0.173
	N2	−0.149 ± 0.111	**−0.121** ± 0.124	−0.245 ± 0.139	−0.360 ± 0.172	**−0.341** ± 0.166	−0.402 ± 0.167
	N3	−0.445 ± 0.100	**−0.402** ± 0.091	−0.485 ± 0.100	**−0.652** ± 0.170	−0.668 ± 0.172	−0.828 ± 0.175
	N4	−0.191 ± 0.127	**−0.134** ± 0.115	−0.824 ± 0.285	**−0.577** ± 0.198	−0.639 ± 0.198	−1.728 ± 0.393
FinCast	N1	−0.010 ± 0.039	**+0.009** ± 0.027	−0.026 ± 0.043	**−0.006** ± 0.031	−0.020 ± 0.022	−0.010 ± 0.033
	N2	+0.045 ± 0.041	**+0.057** ± 0.034	+0.054 ± 0.046	+0.009 ± 0.037	−0.014 ± 0.031	**+0.014** ± 0.040
	N3	−0.119 ± 0.048	**−0.080** ± 0.035	−0.168 ± 0.059	−0.054 ± 0.041	**−0.020** ± 0.025	−0.069 ± 0.047
	N4	+0.018 ± 0.045	**+0.037** ± 0.036	−0.130 ± 0.065	−0.045 ± 0.046	**−0.025** ± 0.031	−0.128 ± 0.077

128 independent scored contexts per rung; mirrors form the correction and are not additional scored samples. Requested future length is 128. Sign avg.: four independent sign draws. N3 uses $\nu = 5$; N4 is the dependent control and does not have independent-sign symmetry. ±: bootstrap SE over series. Bold marks the highest skill among the three forecasts within each row and horizon.

equity windows, so the losses there are not a sampling effect. The part can be negative when the mirrored contexts have larger sample variance than the originals.

G.6 Risk Terms behind Raw-Window Losses

Let $Y = y_{t+h}/y_t - 1$. The empirical skill change splits exactly into $\mathbb{E}[a^2] - 2\mathbb{E}[Ya] + 2\mathbb{E}[(D-a)a]$, divided by the persistence squared error (Table 26), reported as removed energy, future alignment and the cross term.

Alignment and cross terms explain the losses beyond the removed energy. For Chronos-T5 on equities, the three terms are $+0.040$, -0.102 and -0.004; Time-MoE has a cross term of -0.574. Moirai-2.0 is the only equity model to improve at h=128: $+0.059 - 0.081 + 0.058 = +0.036$.

The annualised mean even departure ranges from $+1.5\%$ to $+7.0\%$ for Chronos-T5, Chronos-Bolt, TiRex, Moirai, and FinCast on equities, compared with a realised annualised move of $+13.6\%$ ($+1.8\%$ on FX). Half the mean removed component is not a break-even drift, because a variable even

Table 23: Daily forecasts before and after correction at $h = 128$.

	Mean departure / σ_r			Skill		
Model	**Raw**	**Mirror**	**Sign avg.**	**Raw**	**Mirror**	**Sign avg.**
Sign-randomised daily windows (1280 originals)						
Chronos-T5	+1.54 ± 0.09	+0.03 ± 0.05	+0.00 ± 0.00	−0.072 ± 0.034	**−0.030** ± 0.013	−0.068 ± 0.029
Chronos-Bolt	+1.46 ± 0.10	−0.02 ± 0.04	−0.00 ± 0.00	−0.083 ± 0.032	**−0.038** ± 0.013	−0.102 ± 0.038
Chronos-2	+0.16 ± 0.07	−0.01 ± 0.03	−0.00 ± 0.00	−0.091 ± 0.033	**−0.067** ± 0.025	−0.125 ± 0.042
TiRex	+1.89 ± 0.08	+0.03 ± 0.05	−0.00 ± 0.00	−0.103 ± 0.042	**−0.059** ± 0.024	−0.110 ± 0.038
Moirai-1.1	+2.28 ± 0.14	+0.02 ± 0.06	+0.00 ± 0.00	−0.201 ± 0.096	**−0.080** ± 0.033	−0.221 ± 0.094
Moirai-2.0	−0.12 ± 0.11	−0.11 ± 0.04	−0.00 ± 0.00	−0.116 ± 0.036	**−0.074** ± 0.024	−0.152 ± 0.047
TimesFM-2.0	+0.48 ± 0.11	−0.03 ± 0.07	+0.00 ± 0.00	−0.265 ± 0.097	**−0.152** ± 0.059	−0.322 ± 0.124
TimesFM-2.5	+0.16 ± 0.08	+0.05 ± 0.05	−0.00 ± 0.00	−0.089 ± 0.041	**−0.067** ± 0.031	−0.117 ± 0.046
Time-MoE	+1.96 ± 0.26	+0.16 ± 0.14	−0.00 ± 0.00	−0.945 ± 0.401	**−0.820** ± 0.339	−1.106 ± 0.425
Sundial	+1.74 ± 0.20	+0.10 ± 0.10	−0.00 ± 0.00	−0.516 ± 0.223	**−0.423** ± 0.167	−0.593 ± 0.217
FinCast	−0.26 ± 0.06	+0.05 ± 0.02	+0.00 ± 0.00	−0.053 ± 0.022	**−0.033** ± 0.013	−0.065 ± 0.023
Raw daily windows (1280)						
Chronos-T5	+2.17 ± 0.21	+0.64 ± 0.21	+0.64 ± 0.22	**+0.041** ± 0.058	−0.023 ± 0.035	−0.061 ± 0.025
Chronos-Bolt	+3.13 ± 0.34	+1.08 ± 0.22	+1.67 ± 0.36	**+0.122** ± 0.106	+0.050 ± 0.058	+0.047 ± 0.105
Chronos-2	+1.35 ± 0.25	+1.82 ± 0.41	+1.19 ± 0.27	**−0.023** ± 0.049	−0.038 ± 0.064	−0.063 ± 0.053
TiRex	+1.76 ± 0.34	+0.11 ± 0.32	−0.12 ± 0.35	**+0.115** ± 0.098	−0.073 ± 0.070	−0.009 ± 0.073
Moirai-1.1	+2.38 ± 0.25	+0.12 ± 0.20	+0.10 ± 0.33	**−0.023** ± 0.075	−0.073 ± 0.037	−0.207 ± 0.042
Moirai-2.0	+1.41 ± 0.41	+0.76 ± 0.23	+1.53 ± 0.37	−0.009 ± 0.089	**+0.026** ± 0.052	−0.057 ± 0.082
TimesFM-2.0	−0.15 ± 0.63	−0.77 ± 0.73	−0.63 ± 0.58	**−0.188** ± 0.179	−0.310 ± 0.261	−0.357 ± 0.173
TimesFM-2.5	+0.47 ± 0.32	+0.37 ± 0.31	+0.31 ± 0.34	**−0.014** ± 0.058	−0.024 ± 0.054	−0.087 ± 0.043
Time-MoE	−4.33 ± 0.94	−6.06 ± 0.97	−6.29 ± 0.89	**−0.889** ± 0.288	−1.589 ± 0.487	−1.618 ± 0.318
Sundial	−3.16 ± 0.63	−4.69 ± 0.61	−4.90 ± 0.58	**−0.541** ± 0.200	−0.952 ± 0.296	−1.007 ± 0.234
FinCast	−0.34 ± 0.18	−0.57 ± 0.14	−0.08 ± 0.18	**−0.004** ± 0.030	−0.056 ± 0.025	−0.010 ± 0.034

1280 original windows (380 FX, 900 equities), context 512, four sign copies per window. Mean departure uses σ_r units; skill pools squared return errors. Mirror: one mirror per input. Sign avg.: four draws on raw windows; leave-one-out over sign copies on the null. ±: SE from 37 family-by-future-end-date bootstrap clusters. Bold marks the highest skill among the three forecasts within each row. Subsets are reported in Table 24.

part also contributes its variance, its covariance with the drift and the odd–even cross moment (Equation 5). Historical-mean comparisons in equity forecasting also motivate a drift baseline (Welch & Goyal, 2008; Campbell & Thompson, 2008).

G.7 Shrinkage and the Estimated-Drift Heuristic

A context-dependent gate combines a forecast with persistence:

$$\hat{y}^{\pi} = (1 - \pi)y_t + \pi\hat{y}, \qquad \pi = \pi(x) \in [0, 1]. \tag{6}$$

Under Definition 1, its risk is $\mathcal{R}(y_t) + \mathbb{E}[\pi^2(\hat{y} - y_t)^2]$, which cannot exceed the model's risk (Proposition 3); the bound comes from the null and the shrinkage target rather than from an estimated combination weight (Bates & Granger, 1969; Meese & Rogoff, 1983; Wright, 2008). Candidate context statistics for π include the variance ratio (Lo & MacKinlay, 1988), increment spectral flatness and the Ljung–Box statistic (Appendix A.4). Scalar shrinkage damps both odd and even parts, whereas mirror averaging removes only the even part and keeps an odd conditional mean under the symmetry of N4.

The estimated-drift gate. This separate rule chooses between raw and mirror-corrected forecasts without a risk guarantee on raw windows. It retains the raw forecast when $a(2h\hat{\mu} - a) > 0$. For positive a this corresponds to $h\hat{\mu} > a/2$; for negative a the inequality reverses. It is a constant-benchmark heuristic that omits predictor variation, cross terms and uncertainty in $\hat{\mu}$. With the context mean return as $\hat{\mu}$, the reported application rates are 36–60 percent at h=128 (Table 27). On equities it outperforms the unconditional mirror for ten models, but the raw forecast for only two. Chronos-T5's skill is $+0.048$ raw, -0.018 corrected and $+0.018$ gated. On FX the gate beats the raw forecast for seven models and the mirror for five. **The gate is therefore not a reliable way to choose between the two forecasts.**

G.8 Mirror Augmentation during Fine-Tuning

The mirror correction doubles the cost of inference and leaves the weights as they are. To test whether the symmetry can be trained into a model instead, we fine-tuned the released Chronos-Bolt-small twice from the same weights and on the same windows. The data are the 25 training sets of the

Table 24: Daily correction results by data subset at $h = 128$.

	Mean departure / σ_r			Skill		
Model	**Raw**	**Mirror**	**Sign avg.**	**Raw**	**Mirror**	**Sign avg.**
Raw FX (380 windows)						
Chronos-T5	+1.06 ± 0.21	−0.51 ± 0.22	−0.65 ± 0.29	**−0.172** ± 0.104	−0.174 ± 0.051	−0.313 ± 0.083
Chronos-Bolt	+1.71 ± 0.26	−0.10 ± 0.21	+0.06 ± 0.26	−0.091 ± 0.084	**−0.069** ± 0.040	−0.180 ± 0.075
Chronos-2	−0.03 ± 0.22	−0.41 ± 0.21	−0.42 ± 0.22	−0.127 ± 0.045	**−0.111** ± 0.048	−0.185 ± 0.046
TiRex	+0.76 ± 0.58	−1.17 ± 0.52	−1.35 ± 0.55	**−0.134** ± 0.099	−0.183 ± 0.089	−0.249 ± 0.110
Moirai-1.1	+1.75 ± 0.25	−0.43 ± 0.29	−0.42 ± 0.26	−0.230 ± 0.108	**−0.134** ± 0.056	−0.311 ± 0.099
Moirai-2.0	−0.17 ± 0.51	−0.45 ± 0.33	−0.13 ± 0.53	−0.173 ± 0.076	**−0.149** ± 0.048	−0.244 ± 0.080
TimesFM-2.0	−1.70 ± 0.65	−1.62 ± 0.50	−1.78 ± 0.69	−0.260 ± 0.100	**−0.189** ± 0.058	−0.389 ± 0.113
TimesFM-2.5	−1.31 ± 0.42	−1.27 ± 0.40	−1.41 ± 0.49	**−0.152** ± 0.057	−0.155 ± 0.050	−0.224 ± 0.075
Time-MoE	−2.45 ± 1.05	−2.84 ± 1.07	−3.26 ± 1.07	**−0.809** ± 0.253	−0.877 ± 0.252	−1.150 ± 0.310
Sundial	−1.51 ± 0.86	−2.19 ± 0.87	−2.48 ± 0.88	**−0.511** ± 0.186	−0.581 ± 0.211	−0.695 ± 0.200
FinCast	−1.00 ± 0.19	−0.58 ± 0.22	−0.60 ± 0.20	−0.101 ± 0.041	**−0.045** ± 0.031	−0.087 ± 0.040
Raw equities (900 windows)						
Chronos-T5	+2.65 ± 0.24	+1.13 ± 0.25	+1.18 ± 0.22	**+0.048** ± 0.058	−0.018 ± 0.034	−0.053 ± 0.026
Chronos-Bolt	+3.72 ± 0.42	+1.58 ± 0.25	+2.35 ± 0.44	**+0.129** ± 0.111	+0.054 ± 0.058	+0.054 ± 0.106
Chronos-2	+1.93 ± 0.30	+2.76 ± 0.45	+1.87 ± 0.30	**−0.020** ± 0.052	−0.036 ± 0.066	−0.059 ± 0.054
TiRex	+2.19 ± 0.39	+0.65 ± 0.35	+0.39 ± 0.40	**+0.123** ± 0.099	−0.070 ± 0.072	−0.002 ± 0.073
Moirai-1.1	+2.64 ± 0.32	+0.35 ± 0.24	+0.32 ± 0.44	**−0.016** ± 0.075	−0.071 ± 0.037	−0.204 ± 0.043
Moirai-2.0	+2.08 ± 0.51	+1.27 ± 0.25	+2.23 ± 0.41	−0.004 ± 0.089	**+0.032** ± 0.051	−0.051 ± 0.084
TimesFM-2.0	+0.50 ± 0.82	−0.40 ± 1.05	−0.15 ± 0.76	**−0.186** ± 0.191	−0.314 ± 0.262	−0.356 ± 0.181
TimesFM-2.5	+1.22 ± 0.35	+1.07 ± 0.32	+1.04 ± 0.36	**−0.009** ± 0.059	−0.019 ± 0.055	−0.083 ± 0.045
Time-MoE	−5.13 ± 1.25	−7.42 ± 1.18	−7.57 ± 1.08	**−0.891** ± 0.293	−1.611 ± 0.514	−1.632 ± 0.341
Sundial	−3.85 ± 0.78	−5.74 ± 0.74	−5.92 ± 0.64	**−0.542** ± 0.205	−0.964 ± 0.307	−1.016 ± 0.236
FinCast	−0.07 ± 0.24	−0.57 ± 0.18	+0.13 ± 0.24	**−0.001** ± 0.031	−0.056 ± 0.025	−0.008 ± 0.035
Raw post-2025 subset (280 windows)						
Chronos-T5	+2.29 ± 0.33	+0.86 ± 0.31	+0.78 ± 0.37	**+0.074** ± 0.045	+0.011 ± 0.043	−0.031 ± 0.049
Chronos-Bolt	+2.98 ± 0.52	+0.94 ± 0.37	+1.18 ± 0.48	**−0.010** ± 0.093	−0.018 ± 0.072	−0.075 ± 0.078
Chronos-2	+1.70 ± 0.32	+2.47 ± 0.55	+1.58 ± 0.38	+0.070 ± 0.052	−0.089 ± 0.196	**+0.090** ± 0.095
TiRex	+1.38 ± 0.51	+0.01 ± 0.43	−0.43 ± 0.58	**+0.019** ± 0.052	−0.175 ± 0.119	−0.102 ± 0.050
Moirai-1.1	+2.58 ± 0.42	+0.25 ± 0.26	+0.39 ± 0.58	−0.133 ± 0.082	**−0.100** ± 0.069	−0.215 ± 0.059
Moirai-2.0	+1.63 ± 0.35	+0.93 ± 0.27	+1.80 ± 0.33	+0.023 ± 0.033	**+0.039** ± 0.063	−0.031 ± 0.059
TimesFM-2.0	+0.52 ± 0.35	+0.65 ± 0.42	+0.36 ± 0.34	**−0.159** ± 0.047	−0.215 ± 0.055	−0.222 ± 0.054
TimesFM-2.5	−0.05 ± 0.30	+0.13 ± 0.36	−0.02 ± 0.30	**−0.096** ± 0.050	−0.172 ± 0.067	−0.172 ± 0.052
Time-MoE	−6.88 ± 1.01	−8.30 ± 1.21	−8.08 ± 1.23	**−1.433** ± 0.489	−2.050 ± 0.709	−2.179 ± 0.524
Sundial	−4.85 ± 0.71	−6.35 ± 0.95	−5.95 ± 0.86	**−0.871** ± 0.208	−1.320 ± 0.335	−1.282 ± 0.216
FinCast	−1.29 ± 0.16	−1.13 ± 0.15	−0.91 ± 0.16	**−0.156** ± 0.034	−0.174 ± 0.050	−0.178 ± 0.035

Same readouts, units, correction methods and bolding as Table 23. ±: cluster-bootstrap SE; FX has 19 calendar clusters and the post-2025 subset has 8. Post-2025 windows overlap the family panels. Repeated raw estimates use the correction analysis's bootstrap draws, so their SEs can differ slightly from Table 14; point estimates are unchanged.

Table 25: Observed skill change from the mirror correction and its sampling-noise part.

	N1, h = 16		N1, h = 128		N4, h = 128		Daily null, h = 128		Raw equities, h = 128	
Model	**Change**	**MC**	**Change**	**MC**	**Change**	**MC**	**Change**	**MC**	**Change**	**MC**
Chronos-T5	+0.071	+0.006	+0.154	+0.008	+0.613	+0.002	+0.041	+0.003	−0.066	0.000
Moirai-1.1	+0.010	+0.010	+0.109	+0.023	+0.543	+0.115	+0.121	+0.016	−0.055	+0.004
Sundial	+0.050	+0.002	+0.022	+0.001	−0.062	+0.001	+0.093	+0.001	−0.422	0.000

Change: skill after minus skill before the mirror correction. MC: the part due to averaging two independent sample means, $(\mathbb{E}[v(x)] - \frac{1}{4}\mathbb{E}[v(x) + v(\bar{x})])/S$ divided by the persistence risk, from the stored within-run variance estimates. Ladder columns use the 128 scored draws per rung; daily columns use the 1280 sign-randomised windows (four copies each) and the 900 raw equity windows. Only the three models whose point forecast is a sample mean are shown; for the others the part is zero.

public Chronos corpus other than exchange rates, capped at 20000 series each (136890 series). A window is a dataset, a series and an end point drawn uniformly, with up to 512 context points and a 64-point target, Bolt's native length. Windows that Bolt's mean-absolute scaling cannot handle, with a flat context or a target above fifty times the context's mean absolute value in either orientation, are redrawn. The plain arm trains on the windows as drawn. The augmented arm replaces each window, context and target together, by its mirror $2w_0 - w$ about its first value with probability one half, using a separate random stream, so both arms see the same windows. Both use Bolt's quantile loss, 1500 steps of 64 windows, AdamW at learning rate 10^{-5} with 100 warm-up steps, cosine decay and gradient clipping at one, and three seeds. Every checkpoint is evaluated with the protocol of the main experiments. The released model's direction, ETTh1 and daily values equal those in Sections 3–6; the N1 and N4 skills use a new sample of 512 series at H=16.

Table 26: Terms in the skill change from mirror correction at $h = 128$.

Model	Δ skill	Removed energy	Future alignment	Cross term	Mean removed (% per year)
FX (380 windows)					
Chronos-T5	−0.002 ± 0.063	+0.082	**−0.085**	+0.001	+1.7%
Chronos-Bolt	+0.022 ± 0.068	**+0.087**	−0.057	−0.008	+1.8%
Chronos-2	+0.016 ± 0.021	**+0.012**	+0.009	−0.005	+0.3%
TiRex	−0.048 ± 0.063	+0.072	**−0.079**	−0.041	+2.0%
Moirai-1.1	+0.096 ± 0.088	**+0.157**	−0.058	−0.002	+2.1%
Moirai-2.0	+0.024 ± 0.062	**+0.104**	−0.051	−0.030	+0.2%
TimesFM-2.0	+0.071 ± 0.047	**+0.077**	−0.048	+0.042	+0.1%
TimesFM-2.5	−0.004 ± 0.019	**+0.010**	−0.006	−0.007	−0.0%
Time-MoE	−0.068 ± 0.049	+0.034	−0.040	**−0.061**	+0.5%
Sundial	−0.070 ± 0.051	+0.032	−0.015	**−0.087**	+0.8%
FinCast	+0.056 ± 0.023	+0.025	**+0.038**	−0.007	−0.4%
Equities (900 windows)					
Chronos-T5	−0.066 ± 0.033	+0.040	**−0.102**	−0.004	+4.6%
Chronos-Bolt	−0.075 ± 0.055	+0.061	**−0.200**	+0.064	+6.4%
Chronos-2	−0.016 ± 0.051	+0.054	+0.049	**−0.118**	−2.5%
TiRex	−0.193 ± 0.055	+0.069	**−0.222**	−0.040	+4.9%
Moirai-1.1	−0.055 ± 0.080	+0.112	**−0.176**	+0.009	+7.0%
Moirai-2.0	+0.036 ± 0.049	+0.059	**−0.081**	+0.058	+2.3%
TimesFM-2.0	−0.128 ± 0.104	+0.134	−0.094	**−0.168**	+2.5%
TimesFM-2.5	−0.010 ± 0.021	**+0.026**	−0.024	−0.012	+0.4%
Time-MoE	−0.720 ± 0.299	+0.213	−0.359	**−0.574**	+8.7%
Sundial	−0.422 ± 0.139	+0.105	**−0.282**	−0.246	+6.1%
FinCast	−0.055 ± 0.022	+0.025	**−0.067**	−0.013	+1.5%

Δ skill is mirror minus raw, with a cluster-bootstrap SE. With $Y = y_{t+h}/y_t - 1$ and $P = \mathbb{E}[Y^2]$, the three terms are $\mathbb{E}[a^2]/P$ (removed energy), $-2\mathbb{E}[Ya]/P$ (future alignment) and $2\mathbb{E}[(D-a)a]/P$ (cross term). Their sum equals Δ skill before rounding. Mean removed is $100(252/128)\mathbb{E}[a]$. Bold marks the largest of the three terms in magnitude in each row.

Table 27: Raw, mirror-corrected and gated daily skill at $h = 128$.

	All windows			FX			Equities			Correction fraction
Model	Raw	Mirror	Gated	Raw	Mirror	Gated	Raw	Mirror	Gated	
Chronos-T5	**+0.041**	−0.023	+0.011	**−0.172**	−0.174	−0.218	**+0.048**	−0.018	+0.018	0.40
Chronos-Bolt	**+0.122**	+0.050	+0.064	−0.091	**−0.069**	−0.086	**+0.129**	+0.054	+0.068	0.36
Chronos-2	**−0.023**	−0.038	−0.051	−0.127	**−0.111**	−0.129	**−0.020**	−0.036	−0.049	0.60
TiRex	**+0.115**	−0.073	+0.010	**−0.134**	−0.183	−0.166	**+0.123**	−0.070	+0.015	0.41
Moirai-1.1	**−0.023**	−0.073	−0.074	−0.230	**−0.134**	−0.207	**−0.016**	−0.071	−0.070	0.39
Moirai-2.0	−0.009	+0.026	**+0.028**	−0.173	−0.149	**−0.148**	−0.004	+0.032	**+0.034**	0.47
TimesFM-2.0	**−0.188**	−0.310	−0.255	−0.260	**−0.189**	−0.210	**−0.186**	−0.314	−0.257	0.54
TimesFM-2.5	**−0.014**	−0.024	**−0.014**	**−0.152**	−0.155	**−0.152**	**−0.009**	−0.019	−0.010	0.52
Time-MoE	**−0.889**	−1.589	−1.127	−0.809	−0.877	**−0.765**	**−0.891**	−1.611	−1.138	0.42
Sundial	**−0.541**	−0.952	−0.604	−0.511	−0.581	**−0.484**	**−0.542**	−0.964	−0.608	0.41
FinCast	−0.004	−0.056	**+0.005**	−0.101	**−0.045**	−0.057	−0.001	−0.056	**+0.007**	0.45

Skill pools squared errors in return units, with 1280 windows overall (380 FX, 900 equities). Gated forecasts follow Algorithm 1, using the context mean return for $\hat{\mu}$. Correction fraction is the proportion of all windows on which the gate applies mirror correction. Bold marks the highest skill within each model and subset.

Before the runs we wrote down three predictions: the augmented model's upward fraction on N1 at h=128 is closer to one half than the released and plainly fine-tuned models'; its ETTh1 skill at h=16 is within 0.02 of the plain model's; and its raw-equity skill at h=128 is lower than the plain model's. **All three hold for each of the three seed pairs** (Table 28). Plain fine-tuning moves the direction a little, from 79 to 71 percent upward; augmentation takes it to 58 percent on N1 and to 49 percent on the sign-randomised daily copies, and raises ETTh1 skill slightly, from 0.697 to 0.708. On raw equities the augmented model keeps +0.064 of skill, about what the inference-time correction leaves of the plainly fine-tuned model (+0.060), and its even part there averages −1.7 percent a year against +4.7. Two results are less favourable. The augmented model's loss on the randomised daily copies is unchanged (−0.074 against −0.073), so by Proposition 1 its departures there became balanced in sign without becoming smaller; and its N4 capture is 0.540 against 0.572. The test covers one small model, one fine-tuning budget and three seeds.

Table 28: Chronos-Bolt-small fine-tuned with and without mirror augmentation.

Probe	Released	Fine-tuned	Mirror-augmented
N1 upward fraction, h=128	0.79	0.71 ± 0.03	0.58 ± 0.04
N1 mean departure (σ), h=128	+2.25	+1.78 ± 0.20	+0.29 ± 0.34
N1 skill, h=16	−0.171	−0.121 ± 0.015	**−0.094** ± 0.010
N4 capture of optimum, h=1	0.538	**0.572** ± 0.007	0.540 ± 0.005
ETTh1 skill, h=16	+0.697	+0.697 ± 0.002	**+0.708** ± 0.003
Raw equities, skill, h=128	**+0.129**	+0.126 ± 0.018	+0.064 ± 0.020
Raw equities, mirror-corrected	+0.054	**+0.060** ± 0.014	+0.028 ± 0.007
Raw equities, even part (%/yr)	+6.4	+4.7 ± 0.5	−1.7 ± 1.6
Sign-randomised skill, h=128	−0.083	**−0.073** ± 0.006	−0.074 ± 0.020
Sign-randomised upward fraction	0.71	0.68 ± 0.02	0.49 ± 0.04

Fine-tuned columns: mean ± standard deviation over three seeds; the two arms share initial weights, windows and seeds and differ only in the mirror coin. Direction uses the N1 contexts of Section 3; N1 skill and N4 share use 512 new series at H=16 (N4 share relative to the exact conditional mean). ETTh1 skill is in context-σ units; daily rows use the 900 raw equity windows and the 1280 sign-randomised windows of Appendix F.1, in return units. Mirror-corrected: the inference-time correction applied to each checkpoint. Even part: $100(252/128)\mathbb{E}[a]$ at h=128. Bold marks the highest entry in each skill or share row; upward fractions, mean departures and even parts are not ranked.

H Scope and Unresolved Attribution

Remark 4 (Scope of the null evaluation) *Null scores measure behaviour under a fixed conditional mean. A model with large excess risk there can still be useful on predictable series. N4, ETTh1 and the small-cap comparisons test specific alternatives and rank the models only on those.*

Attribution. The null measures what a departure costs, not where it comes from: context-fitted models also depart under a known null (Andreoletti, 2026), frozen weights can respond to chance patterns, and outside the assumptions of Proposition 2 an even component can be useful or harmful. The controlled corpus comparison identifies the intervention effect of Section 4. Beyond the trend statistics of Appendix E.4 and the search for the market in Appendix F.5, we have not audited financial pretraining subsets; benchmark overlap (Meyer et al., 2025) is addressed only through the raw-data time split (Appendix F.1). Stale prices and bid–ask bounce produce return autocorrelation (Lo & MacKinlay, 1990; Roll, 1984), and small-cap drift is large enough that model gains there cannot be attributed to autocorrelation alone.

Liquidity and market scope. Differences between the smallest and largest size portfolios combine drift, volatility, dependence and portfolio construction, so the relation between liquidity and foundation-model skill remains open.

I Implications and Limitations

Implications. Three checks follow for anyone evaluating a pretrained forecaster on a series close to a martingale. First, report the last value and a drift fitted to the context as baselines, and the share of rising windows for directional accuracy (Table 21); on the small-cap decile no model beat the fitted drift by more than two standard errors, and on rising equities none beat always forecasting a rise. Second, score the model on sign-randomised copies of the same windows, which keep the magnitudes and remove any predictable mean; in our data every model lost on them. Historical skill that times the market from the history but not from the copies or mirrors may be remembered; public training data can be searched for it. Third, report how much of any skill survives removing the even part of the departure, and use Proposition 2(iii) to attribute what is lost; on rising equities most of it did not survive. The even-energy ratio measures how symmetric a forecast is, not whether its asymmetry is useful, and on series with a real drift either form of the correction should be checked before use.

Limitations. The null processes form a small synthetic family, and real returns satisfy the null only approximately. N4, ETTh1 and the small-cap decile show that the models can use structure, not that they beat fitted baselines, and ETTh1 is a public benchmark that some pretraining corpora may contain. Point outputs differ across models, and for Moirai-1.1 part of the mirror's gain is reduced sampling noise. The corpus experiment uses small models, of 0.54M and 3.18M parameters, and three seeds; the corpus audit covers three public corpora whose training weights we do not know,

and the fine-tuning test covers one model. Public training data contain the remembered years, but Chronos-Bolt does not document its training data, and the calendar-shift check is clear only for TimesFM-2.0. The daily standard errors miss some common shocks, the transaction data cover one day, and we have not tested a liquidity mechanism (Appendix H).

J Related Work

Time series foundation models. The released models differ in how they encode values and what they are trained to output. Chronos (Ansari et al., 2024) quantises scaled values into tokens for a language-model backbone and Chronos-2 (Ansari et al., 2025) encodes patches; Moirai (Woo et al., 2024) is a masked encoder over patches; TiRex (Auer et al., 2025) is a recurrent xLSTM; TimesFM (Das et al., 2024), Time-MoE (Shi et al., 2025), Sundial (Liu et al., 2025b), Moirai-2.0 (Liu et al., 2025a) and FinCast (Zhu et al., 2025) are decoder-only. Their losses include token cross-entropy, quantile loss, mixture likelihoods, Huber loss and flow matching (Appendix B.2).

Evaluations against simple baselines. On financial data, Rahimikia et al. (2025) report negative out-of-sample R^2 for a model pretrained on securities' daily returns and limited gains from fine-tuning, and Noguer i Alonso & Pereira Franklin (2026) find little improvement over a random walk on liquid US equities; the finance-native model we test was released without a persistence baseline (Zhu et al., 2025). Synthetic probes intervene on generator parameters (Jander et al., 2026) or use random walks to ask which summary of the future a loss returns (Yu et al., 2026), and Andreoletti (2026) uses a decomposition related to Proposition 1, with fitted noise as the excess. Studies on real returns compare models where the optimum is unknown; we measure against a known optimum and ask what a model adds, which is what lets us separate a directional bias from skill. Pretraining data can also contain the evaluation series, or correlated series from the same period (Meyer et al., 2025). Li et al. (2026) audit such contamination per dataset from how quickly a model adapts when fine-tuned on it, and language models reproduce economic series from before their training cut-off but not after it (Lopez-Lira et al., 2025); our checks for remembered history use forecasts and, where the training data are public, the data themselves.

Predictable components of returns. Where liquid returns are predictable, the cause is often drift or microstructure: stale prices induce positive autocorrelation in indices and portfolios (Lo & MacKinlay, 1990), and bid–ask bounce induces negative first-order autocorrelation in individual securities (Roll, 1984), on which our positive control is built. Such predictability need not survive trading costs.

Other forecasters and benchmarks. GIFT-Eval (Aksu et al., 2024) and long-horizon suites (Wu et al., 2023; Liu et al., 2024a) evaluate forecasters with other baselines and tasks. TimeGPT has closed weights and an API interface (Garza et al., 2023), MOMENT fits a forecasting head for the downstream task (Goswami et al., 2024), and Lag-Llama (Rasul et al., 2023) is a further open forecaster; the probe applies to any forecaster that returns a point prediction. Work comparing linear and transformer forecasts (Zeng et al., 2023) likewise argues for explicit baselines, which under a martingale null is persistence.

Symmetry and forecast combination. The correction is standard orbit averaging (Chen et al., 2020; Lyle et al., 2020), and the gates we test combine a forecast with persistence (Bates & Granger, 1969; Timmermann, 2006). Appendix H discusses scope and what the evidence leaves unidentified.