BEYOND FORECAST ACCURACY: HOW TIME-SERIES FOUNDATION MODELS RESPOND TO LONGER HISTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Longer histories can stabilize forecasts without correcting their average response to predictive information. We study this distinction in six frozen time-series foundation model checkpoints on Gaussian processes with known next-step distributions. At 2048 observations, the five deterministic checkpoints respond too weakly to the true predictive mean on average. Their excess mean squared errors are 24–45% of predictable variance, compared with 6.5% for a fitted AR(64) baseline. Yet Chronos-Bolt and Moirai vary less across histories with the same predictive mean and last observation than that baseline. Linear regression exposes the response bias. Conditional resampling then separates two sources of the regression residual: a curved average response and variation across matched histories. Exploratory bounds suggest that most residual energy comes from the latter. History length also changes the association with the last observation at a fixed predictive mean. Following exploration, a prospective three-lag test supports opposite Bolt–Moirai changes from 512 to 2048, averaged over 16 conditions; Moirai's last-observation response varies nonmonotonically with history length. These comparisons show why greater forecast stability need not repair the main sources of error.

1 INTRODUCTION

More history gives a forecasting model more evidence about its process. Lower error, however, does not tell us how the model used that evidence. Predictions can become more consistent across histories while still responding poorly to the correct mean. We study these possibilities in six frozen checkpoints from the Chronos, TimesFM, Moirai and Sundial families (Ansari et al., 2024; Das et al., 2024; Ansari et al., 2025; Liu et al., 2025a;b).

A useful comparison holds the true predictive distribution fixed while changing the history presented to the model. Within each stationary Gaussian process, the next-step mean S determines that distribution. We generate alternative histories with the same S, then also fix the last observation A (Figure 1). The model receives only the history, without the process parameters or S. Different histories can help it estimate those unknown parameters, so variation at fixed (A, S) is not automatically a defect. We study how that variation combines with bias and forecast error.

Stable predictions can still be inaccurate. At length 2048, Chronos-Bolt and Moirai vary less across histories sharing (A, S) than a fitted AR(64) baseline, yet their excess errors are about six to seven times larger. The five deterministic checkpoints have average responses to S below the ideal value of one. Bias and linear response mismatches account for most of their excess error in this experiment. A stability measure alone would miss this weak average response.

Resampling explains what a linear fit leaves behind. The response bias is visible in ordinary regression. But two predictors can have identical linear fits and residual errors while behaving differently: one follows a curved average response; the other varies across histories with the same predictive information. Paired resampling distinguishes them without fitting a nonlinear response function. Exploratory bounds in our long-history confirmation experiment suggest that most of the regression residual comes from variation across histories sharing (A, S).

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

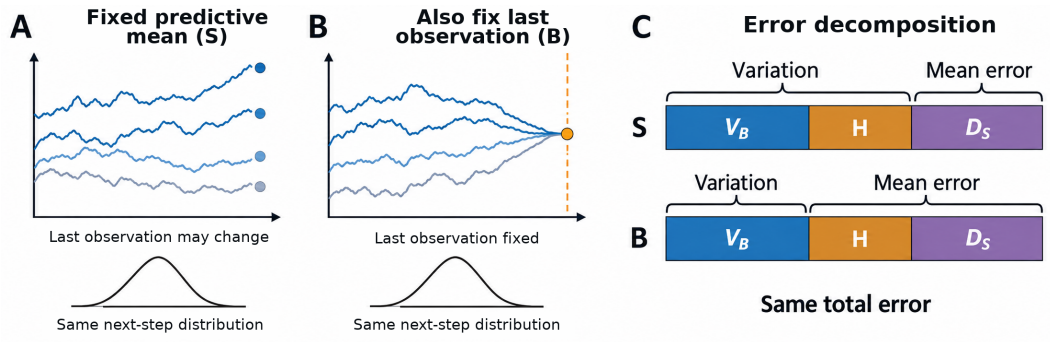

Figure 1: **Resampling histories separates three sources of forecast error.** (A) Hold the predictive mean S fixed. (B) Also hold the last observation A fixed, giving $\mathcal{B} = (A, S)$ (shown as B). Both preserve the next-step distribution. (C) Excess MSE is $\mathcal{E} = V_{\mathcal{B}} + H + D_S$: remaining history variation, the last-observation component, and conditional mean error. The braces show how the same error splits when different information is held fixed. Curves and bar widths are schematic.

Longer histories change more than one response. Two exploratory experiments suggest that extending history from 512 to 2048 weakens Bolt's association with the last observation at fixed S, while strengthening Moirai's. A prospective three-lag experiment supports this average contrast on fresh histories. Moirai's last-observation response varies nonmonotonically with history length: it first falls and then rises; Bolt's decline is concentrated in processes with a smaller local coefficient share. The contrast therefore depends on the measured checkpoint, process and length interval.

Our contribution is to connect these response errors with their conditional sources and changes with history. The comparison builds on established conditional projection, Gaussian simulation and pick–freeze estimation (Goda, 2017; Wilson et al., 2020; Janon et al., 2014).

2 Comparing histories with the same predictive distribution

What we hold fixed. A pipeline receives a chronological history $X = (Y_0, \ldots, Y_{n-1})$ and forecasts Y_n; write its median as $f_n(X)$. Within each stationary Gaussian condition, the correct next-step distribution is $N(S, \sigma_F^2)$, where $S = \mathbb{E}[Y_n \mid X]$ and σ_F^2 is constant. Thus histories with the same predictive mean S have the same predictive distribution. We resample histories holding either S or both S and the last observation $A = X_{n-1}$ fixed. Write $\mathcal{B} = (A, S)$ and call the resulting histories X^S and $X^{\mathcal{B}}$ conditional replicas.

The discovery and confirmation experiments use two predictive lags:

$$Y_t = aY_{t-1} + bY_{t-L} + \varepsilon_t, \qquad S = aA + bP, \quad P = X_{n-L}, \quad \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2). \tag{1}$$

Every condition has unit marginal variance. We set $a = cr$, $b = \pm c(1-r)$, where $c = |a| + |b|$ and r is the last observation's coefficient share. Since $a, b \neq 0$, fixing (A, S) also fixes P; in the three-lag test, other predictive inputs remain free (Section 6). All central history lengths already contain the predictive lags. Longer histories can still help estimate unknown dynamics: the predictor receives neither the coefficients nor S. Variation at fixed (A, S) can therefore reflect uncertainty about these parameters. Appendix A.3 extends the construction to latent sums.

How we resample. For $X \sim N(0, \Sigma)$, let the full-row-rank matrix C select the quantities to hold fixed. With an independent donor $U \sim N(0, \Sigma)$, set

$$\mathsf{G} = \Sigma C^\top (C\Sigma C^\top)^{-1}, \qquad X^{\mathcal{B}} = \mathsf{G}CX + (I - \mathsf{G}C)U. \tag{2}$$

The first term retains the fixed information; the second supplies independent remaining variation (Wilson et al., 2020). Keeping only the row $CX = S$ produces X^S. Each replica follows the original conditional history distribution, preserving population forecast risk while allowing paired predictions to differ. Both replica types share a donor.

Table 1: Scientific questions and how the experiments relate. Two-lag response results motivated the subsequent prespecified three-lag test.

Experiment	Scientific question and source
Two-lag discovery	How do accuracy and response change with history length? Eight conditions supplied exploratory findings.
Two-lag confirmation	Do the predicted error-component changes recur? Response trends were examined after observing this experiment too.
Three-lag test	Do the opposing response trends persist with an additional predictive lag? Both two-lag experiments motivated the prespecified 512–2048 hypotheses tested on fresh histories.

Experiments and uncertainty. Discovery uses eight conditions with $L = 13, 29$ and lengths $128, 512, 2048$; confirmation uses 16 distinct conditions with $L = 19, 41$ and lengths $512, 2048$. Each condition has three seeds with 256 or 128 independent base/donor tuples per seed, respectively. Six checkpoints share the inputs; shorter histories are exact suffixes at the same target time. Averages weight conditions equally. Ten thousand paired bootstrap draws resample tuples within condition and seed, retaining pairing across replicas, models and lengths. Conditions remain fixed: intervals describe trajectory uncertainty within these designs. Sundial also conditions on its recorded prediction stream (Appendix B.1). Full designs and inference details appear in Appendices F and B.

Table 1 tracks the source of each scientific question. Baseline and normalization tests have separate rules; two-lag response, AR(64)-baseline and condition-group analyses are post hoc. Protocols were recorded internally, not publicly preregistered.

Forecast error. Point excess error is $\mathcal{E}_n = \mathbb{E}[(f_n - S)^2]$, excluding innovation noise. Central experiments request and score one step. Supplementary nine-quantile risk (Q9) and requested-horizon controls appear in Appendices G and H.

3 Separating forecast variation, mean error and response

3.1 Three components of forecast error

Average the forecasts over histories sharing (A, S) or S alone: $g_{\mathcal{B},n} = \mathbb{E}[f_n \mid \mathcal{B}]$ and $g_{S,n} = \mathbb{E}[f_n \mid S]$. Variation around the first average gives

$$V_{\mathcal{B},n} = \tfrac{1}{2}\mathbb{E}[(f_n(X^{\mathcal{B}}) - f_n(X))^2] = \mathbb{E}\,\mathrm{Var}(f_n \mid \mathcal{B}), \tag{3}$$

$$\mathcal{E}_n = V_{\mathcal{B},n} + D_{\mathcal{B},n}, \qquad D_{\mathcal{B},n} = \mathbb{E}[(g_{\mathcal{B},n} - S)^2]. \tag{4}$$

Here $V_{\mathcal{B},n}$ measures *remaining history variation* and $D_{\mathcal{B},n}$ error in the average forecast. Comparing the averages gives

$$H_n = \mathbb{E}[(g_{\mathcal{B},n} - g_{S,n})^2], \qquad V_{S,n} = V_{\mathcal{B},n} + H_n, \qquad D_{\mathcal{B},n} = H_n + D_{S,n}, \tag{5}$$

where $D_{S,n} = \mathbb{E}[(g_{S,n} - S)^2]$. The *last-observation component* H_n captures how the average forecast varies with A at fixed S; *conditional mean error* $D_{S,n}$ remains after averaging over all histories sharing S. Their sum with remaining history variation gives total excess error. Appendix A.2 gives the proof.

For errors $e_0, e_{\mathcal{B}}, e_S$ relative to the common S,

$$\widehat{H} = \mathrm{Avg}[e_0(e_{\mathcal{B}} - e_S)], \qquad \widehat{D}_S = \mathrm{Avg}[e_0 e_S] \tag{6}$$

estimate the components without fitting conditional means (Goda, 2017). With the paired squared difference, they exactly decompose the paired total, which can differ in sample from original-forecast error $\widehat{\mathcal{E}}_0 = \mathrm{Avg}[e_0^2]$. Signed estimates are retained (Appendix B).

3.2 Two linear responses, with different meanings

The predictive-mean response β describes forecast amplitude as S changes. The last-observation response α measures the linear association with A after accounting for S. Set $v = \mathrm{Var}(S)$, $\gamma_{AS} =$

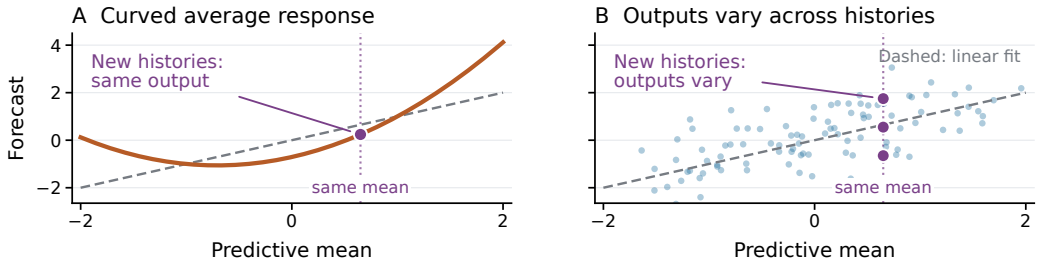

Figure 2: **The same linear fit can hide different behavior across histories.** Analytic examples, not checkpoint results: independent standard Gaussians U and T represent the predictive mean and remaining history. The forecasts are $U+(U^2-1)/\sqrt{2}$ (left) and $U+T$ (right). Both have population linear fit U (dashed) and population residual MSE 1; the dots are illustrative draws, not an empirical fit. Fixing the predictive mean U and resampling histories leaves the left forecast unchanged but changes the right; their history variances are 0 and 1.

$\mathrm{Cov}(A,S)$, $Z_A = A-(\gamma_{AS}/v)S$ and $q = \mathrm{Var}(Z_A) > 0$. Projecting onto the uncorrelated directions S and Z_A gives

$$f_n \longmapsto \eta_n + \beta_n S + \alpha_n Z_A, \qquad \eta_n = \mathbb{E}[f_n - S],$$
$$\beta_n = 1 + \frac{\mathbb{E}[(f_n - S)S]}{v}, \qquad \alpha_n = \frac{\mathbb{E}[(f_n - S)Z_A]}{q}. \tag{7}$$

These are statistical associations, not causal input weights: $f = bP$ never uses A explicitly, yet has $\alpha = -a$ when $S = aA + bP$. The ideal values are $(\beta, \alpha) = (1, 0)$; moving α toward zero need not improve accuracy if β moves away from one. Orthogonality gives

$$H_n^{\mathrm{lin}} = \alpha_n^2 q \le H_n, \qquad D_{S,n} = \underbrace{\eta_n^2 + (\beta_n - 1)^2 v}_{D_{S,n}^{\mathrm{affine}}} + D_{S,n}^{\mathrm{rem}}. \tag{8}$$

The remainders $H - H^{\mathrm{lin}}$ and D_S^{rem} are nonnegative in population. In this two-lag process, fixing (P, S) also fixes (A, S) and gives the same H (Appendix D.4).

What resampling adds to a linear fit. An equally good linear fit can hide two different behaviors (Figure 2). Paired forecasts share the same conditional mean; their deviations around it vary independently. Original forecasts suffice to estimate the slopes and affine residual energy, with $r = f - \eta - \beta S - \alpha Z_A$:

$$R_{\mathrm{aff}} := \mathbb{E}[r^2] = V_{\mathcal{B}} + (H - H^{\mathrm{lin}}) + D_S^{\mathrm{rem}}. \tag{9}$$

The full error is $\mathcal{E} = \eta^2 + (\beta - 1)^2 v + \alpha^2 q + R_{\mathrm{aff}}$. Resampling reveals how the residual divides between remaining history variation and nonlinear conditional means; identical slopes and residual energy can hide different allocations (Appendix D.1).

We use known covariances and independent-seed moment products, recomputed in each bootstrap draw, to avoid bias from squaring noisy estimates. Two 1000-panel simulations assess residual-energy intervals and joint response decisions (Appendices E.1 and E.2).

4 Stable forecasts can still miss the predictive mean

A baseline comparison separates stability from accuracy. At length 2048 in the two-lag confirmation experiment, Bolt and Moirai have less remaining history variation than dense AR(64) least squares (Table 2). Their average forecasts are nevertheless much less accurate. Total excess error is 44.7% of predictable variance for Bolt and 38.6% for Moirai, compared with 6.5% for OLS. All three ratios use the same fixed-condition weights and denominator. Thus, reducing variation across histories would leave much of the two checkpoints' error untouched. These comparisons exclude unpredictable innovation noise. Adding the mean innovation variance gives total MSE 0.80934 for Bolt and 0.78818 for Moirai, versus 0.67744 for OLS: 19.5% and 16.3% higher, respectively. These are descriptive point comparisons; Appendix D.5 gives the calculation.

Table 2: **Smaller history variation does not imply smaller forecast error.** Results at history length 2048 average the 16 fixed confirmation conditions equally. The first three numerical columns use 10^{-3} squared target units; excess error excludes innovation noise. β is the predictive-mean response. Brackets show pointwise 95% paired intervals. The last-observation component H is not shown, and $\mathcal{E}_0$ is estimated separately from original histories. Signed estimates are retained.

Predictor	Excess error $\mathcal{E}_0$	History variation $V_{\mathcal{B}}$	Mean error D_S	Response β
Chronos 2	91.59 [87.74, 95.53]	35.42 [33.78, 37.11]	30.16 [27.10, 33.17]	0.766 [0.749, 0.783]
Chronos-Bolt	154.29 [147.33, 161.26]	18.88 [18.13, 19.66]	102.74 [96.68, 108.77]	0.503 [0.480, 0.526]
TimesFM 3.0	82.94 [79.54, 86.47]	31.29 [29.84, 32.75]	35.26 [32.50, 38.01]	0.697 [0.681, 0.714]
TimesFM 2.5	104.18 [99.55, 108.73]	33.79 [32.34, 35.30]	55.28 [51.54, 59.03]	0.634 [0.615, 0.652]
Moirai 2.0	133.13 [127.00, 139.33]	17.44 [16.69, 18.21]	78.54 [73.49, 83.57]	0.561 [0.539, 0.582]
Dense OLS	22.39 [21.53, 23.28]	22.01 [21.20, 22.86]	0.16 [−0.43, 0.74]	1.002 [0.994, 1.010]
Dense ridge	20.31 [19.54, 21.08]	19.16 [18.47, 19.87]	1.01 [0.49, 1.55]	0.947 [0.940, 0.955]

The average response to predictive information is too weak. The five deterministic checkpoints have mean β between .50 and .77 at 2048, whereas OLS is near one. A response below one means that changes in the correct mean are attenuated in the linear projection of the forecast. This is especially consequential for Bolt and Moirai: D_S is their largest error component. Across the five checkpoints, bias and errors in the two linear responses account for about 62–88% of excess MSE in point estimates. This is an allocation of error using a separate projection in each process, not a single linear model shared by all conditions.

Resampling locates the remaining error. The slopes above need only original forecasts. Figure 3 shows the corresponding empirical separation between remaining history variation and the last-observation and mean-error components. The paired histories answer the next question: what is left after those linear responses are removed? Exploratory upper bounds on the nonlinear conditional-mean share are about 10–11% for Chronos-2 and both TimesFM checkpoints, and 30% and 24% for Bolt and Moirai. Under these bounds, most residual error comes from variation across histories sharing (A, S), rather than curvature in their average forecast. A scalar-calibrated OLS reference reaches the same aggregate history-variation levels with $\lambda = 0.926$ for Bolt and 0.890 for Moirai, yet has excess errors 0.0210 and 0.0218, compared with 0.1543 and 0.1331 for the native checkpoints (Appendix D.1). This is a descriptive comparison of saved forecasts, not a deployable shrinkage rule or a mechanistic explanation. The residual-share conclusion concerns the regression residual; it does not make the full predictor linear. Appendix D.1 gives the estimates and their exploratory, nominal intervals. A separate residual-product check reduces bootstrap variation but shows undercoverage in one simulation case (Appendix D.2).

Fitted baselines improve through a different balance. OLS and ridge are refitted on every history using all 64 lags, without the true period or coefficients; ridge selects its penalty by generalized cross-validation within that history. From 512 to 2048, OLS retains 20% of its history variation and ridge 39%, versus 52–88% for the checkpoints. OLS's mean (β, α) moves from $(1.030, -.022)$ to $(1.002, -.007)$: nonzero variation coexists with an accurate average forecast. These linear processes favor AR baselines. Appendices D.5 and F give further comparisons, including sparse BIC.

5 How more history changes the error

Less variation need not mean a better average response. From 512 to 2048, both Bolt and Moirai become less variable across histories with the same (A, S) in the two-lag confirmation ex-

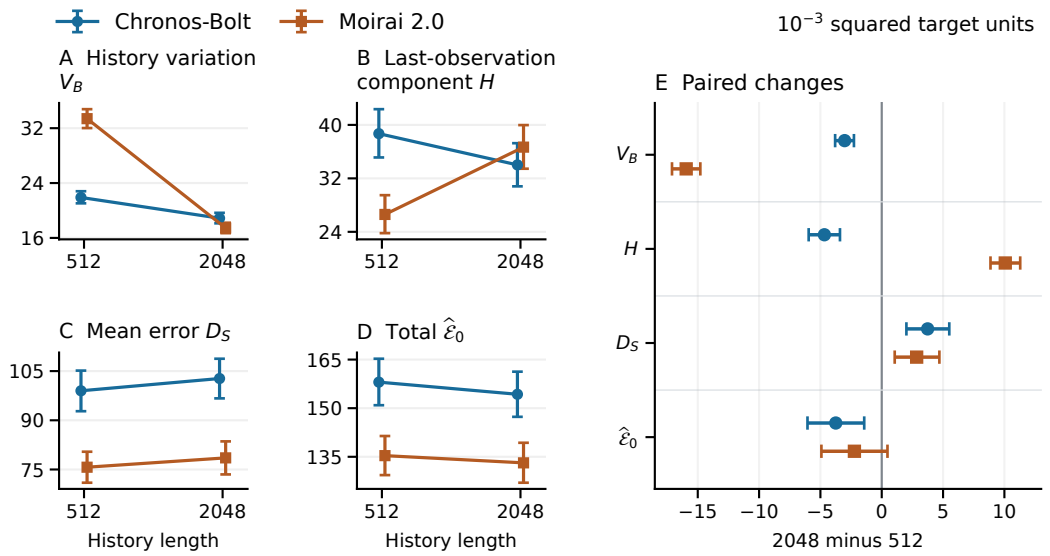

Figure 3: **Both models become less variable across histories, but their last-observation components move in opposite directions.** From 512 to 2048, native Bolt and Moirai reduce remaining history variation V_B and increase conditional mean error D_S; Bolt's last-observation component H falls and Moirai's rises. Equal averages cover all 16 fixed two-lag confirmation conditions. A–D show levels; E shows paired changes; bars are pointwise 95% intervals. Moirai's total-error change interval crosses zero. The total from original histories is estimated separately from the components, so the estimates need not sum exactly.

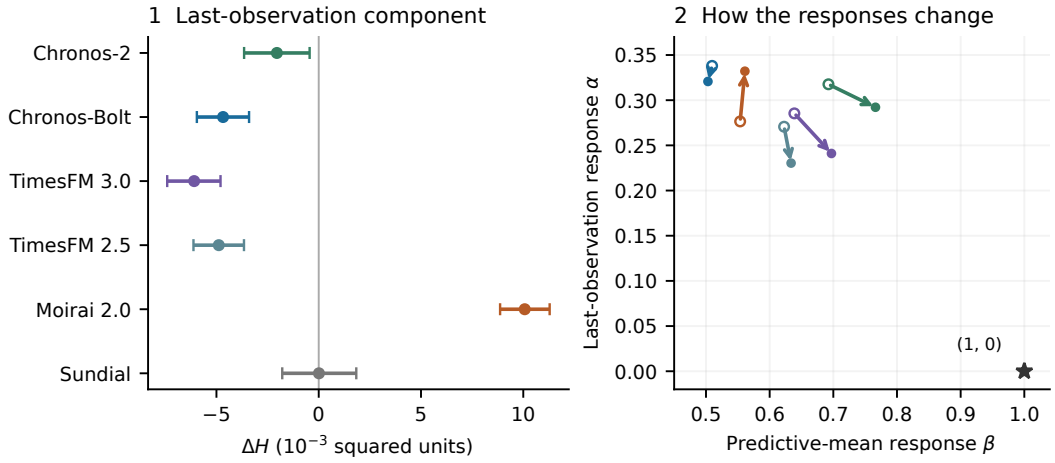

Figure 4: **Longer histories shift the last-observation response differently across checkpoints.** Left: change in the last-observation component H from 512 to 2048, averaged over the 16 fixed confirmation conditions; bars are pointwise 95% paired intervals. Right: post hoc linear responses to the predictive mean (β) and to the last observation after controlling for that mean (α), for the five deterministic pipelines. Open and filled circles mark 512 and 2048. The ideal slopes $(1, 0)$ do not imply zero forecast error. Sundial appears only on the left under its recorded prediction stream.

periment. Yet both increase their conditional mean error D_S (Figure 3). Their responses change in different ways. Bolt's mean β falls by .0065, moving farther from one. Moirai's mean β rises by .0074, but its mean squared slope error increases: its response improves in some conditions and worsens in others. Averaging slopes can therefore conceal errors that remain visible when squared within each process (Appendix D.4).

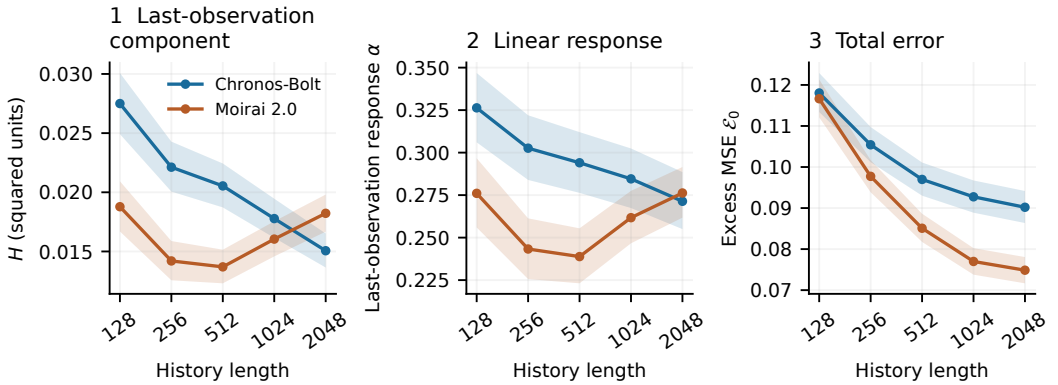

Figure 5: **Moirai's last-observation response rises while its forecast error falls from 512 to 2048.** Results average the 16 fixed three-lag conditions equally. Even after fixing the last observation A and predictive mean S, one combination of the other two predictive lags remains free to vary. The five-length curves show that the last-observation response is not monotone in history length. Shading gives pointwise 95% paired intervals; the prespecified tests compare only 512 and 2048 with adjusted intervals (Table 3). Other lengths are descriptive.

Association with the last observation moves in opposite directions. Let δ denote 2048 minus 512. In both exploratory two-lag comparisons, four deterministic checkpoints reduce mean α, while Moirai increases it. Confirmation gives $\delta\alpha = -.0173$ $[-.0232, -.0111]$ for Bolt and $+.0556$ $[+.0484, +.0626]$ for Moirai (pointwise 95% intervals). Their corresponding last-observation components change by $\delta H = -.00468$ and $+.01007$. Chronos-2 and both TimesFM checkpoints also reduce H; Sundial's interval spans zero under its recorded stream (Figure 4). This association describes the output projection at fixed S, not independent reliance on the last input.

The discovery experiment shows the same 512–2048 directions, but Moirai's response is not monotone across all three lengths. Error components also vary with the process. Of six prespecified component and accuracy predictions in two-lag confirmation, only three are supported, so its overall criterion is not met. In particular, Moirai's total-error decrease remains unresolved. The full family is retained in Table 4. The response contrast was identified from these two experiments and then specified in advance for the new three-lag test below.

The contrast persists under population normalization. Replacing each model's normalization statistics by population center zero and scale one leaves both $\delta\alpha$ signs unchanged. It reduces the Bolt–Moirai difference in δH from $-.01475$ to $-.01307$. Only one of the two prespecified variation reductions is supported; Bolt's overall error and Q9 gains remain unresolved. This modification changes the size of the contrast without identifying its cause (Appendix B).

6 Testing the response contrast with three predictive lags

In a two-lag process, fixing (A, S) fixes both predictive inputs. We next ask whether the Bolt–Moirai response contrast persists when other predictive inputs can still vary. The new process is $Y_t = aY_{t-1} + bY_{t-L_1} + dY_{t-L_2} + \varepsilon_t$. Its 16 conditions vary the two nonlocal lags, coefficient magnitudes and signs. Each supplies three seeds of 128 fresh tuples at nested lengths 128, 256, 512, 1024 and 2048 (Appendix C.1).

Before forecasting, we specified decreasing α for Bolt, increasing α for Moirai, and a negative between-model contrast $J = \delta H^{\text{Bolt}} - \delta H^{\text{Moirai}}$, all from 512 to 2048. All three predictions are supported by the jointly adjusted intervals (Table 3). Moirai's mean α rises by .0374 and Bolt's falls by .0227. Moirai's H and α rise in all 16 condition point estimates; Bolt's fall in 13 and 12, respectively. Thus the average contrast persists, with variation across conditions.

The last-observation response is not monotone. Moirai's last-observation response falls at shorter lengths and then rises, returning near its initial level at 2048 (Figure 5). The exploratory

Table 3: The three-lag experiment supports all three predicted response trends. Changes compare 2048 with 512 over the 16 fixed conditions; $J = \delta H^{\mathrm{Bolt}} - \delta H^{\mathrm{Moirai}}$ compares the models, and α is the last-observation response after controlling for the predictive mean. The prespecified rule requires all three intervals to exclude zero in the predicted direction. Each two-sided interval uses tails $.05/6$, giving a nominal 95% familywise confidence level.

Predicted direction	Estimate	Adjusted interval	Supported?
$J < 0$	-0.01000	$[-0.01131, -0.00870]$	Supported
$\delta\alpha^{\mathrm{Moirai}} > 0$	0.03741	$[0.02762, 0.04757]$	Supported
$\delta\alpha^{\mathrm{Bolt}} < 0$	-0.02266	$[-0.03186, -0.01378]$	Supported

128–2048 change is .0002 with a pointwise 95% interval spanning zero; the prospective finding concerns only 512–2048. Over that interval, both models' original-forecast excess errors decrease: by .00678 for Bolt and .01022 for Moirai, with descriptive intervals excluding zero. Moirai's D_S also falls here, whereas it rose in two-lag confirmation (Table 7). More history therefore improves different parts of the forecast in different processes.

The local coefficient share helps locate the contrast. An exploratory split by r, the true coefficient share of the last observation, puts Bolt's mean $\delta\alpha$ at $-.0480$ when $r = .3$ and $+.0027$ when $r = .7$. The latter interval spans zero, while the direct between-group difference is positive. Moirai's raw slope change is larger at $r = .7$, but after scaling each condition's response by $\sqrt{q}$, the interval for the group difference includes zero (Appendix C.1.1). These comparisons locate heterogeneity without isolating its cause. Projections onto the other lag directions describe related output associations (Appendix C.1), rather than separate input effects.

7 Discussion

Bolt and Moirai give more stable forecasts than OLS across matched histories at 2048, yet track the correct predictive mean poorly. Regression reveals this weak response. Resampling suggests that most remaining residual energy comes from history variation, under exploratory bounds.

The prospective 512–2048 test supports opposite Bolt–Moirai response changes. Moirai's last-observation response is nonmonotone across the measured lengths, and the condition differences show why this is not a universal model trait. Predicting the contrast on new process conditions is a useful next test.

Relation to existing evidence. Synthetic forecasting benchmarks and controlled architecture comparisons expose failures and inductive biases (Tan et al., 2025; Ghoroghchian et al., 2026; Yu et al., 2026). Yu et al. (2026) relate training losses to forecasts of means or medians. Both are optimal and equal S here; $\beta < 1$ measures further attenuation relative to that correct target. Resampling separates curvature from history variation in the remaining error. It also complements conditional importance methods (Strobl et al., 2008; Watson & Wright, 2021): replacement here preserves population risk, while paired outputs reveal variation hidden by that average.

Scope. These findings concern fixed checkpoints, known Gaussian processes and first-step forecasts. Linear processes favor the AR baselines; differences in checkpoint size, architecture and training prevent a mechanistic comparison. History variation need not be independent noise, and average slopes need not describe every condition. Appendix H found no transferable benefit from a separate phase-permutation intervention. Extension to unknown real-process distributions and multi-step forecasting remains open.

REPRODUCIBILITY STATEMENT

The appendices provide conditional-sampling proofs, experiment designs, checkpoint configuration and uncertainty calculations. The anonymous package includes saved forecasts, predictive means and lagged inputs, analysis code, complete numerical tables, figures and LaTeX sources. Its CPU entrypoint reconstructs the conditional components, response projections, reference comparisons and prespecified decisions with fresh paired bootstraps, including refitted held-out affine comparisons. It also reconstructs calibration summaries from retained independent simulations; optional modes rerun the simulations and regenerate histories to refit references. The nested-history reference snapshot supports point reconstruction and saved-draw checks. Fresh checkpoint inference requires the specified GPU environments and model weights. Appendix controls and a separate worked example document their respective reproduction scope.

AI USE STATEMENT

Generative AI tools supported three aspects of this work: (i) writing assistance and language polishing; (ii) retrieval and discovery of related literature; and (iii) research ideation and execution, including hypothesis refinement, experiment design, implementation of simulation and analysis code, and result interpretation. The conceptual overview figure was created with an AI image-generation tool; numerical result figures are drawn from saved experimental outputs. Time-series observations were generated by sampling the explicitly specified stochastic processes using simulation programs; LLM outputs were not used as time-series observations. The authors take responsibility for the cited sources, methods, analysis code, numerical results and final content of this work.

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. URL https://arxiv.org/abs/2410.10393.

Amazon. Chronos-Bolt (base) model card, 2024. URL https://huggingface.co/amazon/chronos-bolt-base. Accessed 2026-09-05.

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

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. URL https://arxiv.org/abs/2510.15821.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Nafiseh Ghoroghchian, Haipeng Zhang, Shuyi Han, Alex Labach, and George Stein. Foundations without fundamentals: Zero-shot blind spots in time series FMs. In *Second ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://openreview.net/forum?id=iIRdd86Xkr.

Takashi Goda. Computing the variance of a conditional expectation via non-nested Monte Carlo. *Operations Research Letters*, 45(1):63–67, 2017. doi: 10.1016/j.orl.2016.12.002. URL https://arxiv.org/abs/1605.05454.

Google. TimesFM 2.5 200m PyTorch model card, 2025. URL https://huggingface.co/google/timesfm-2.5-200m-pytorch. Accessed 2026-09-05.

Google. TimesFM 3.0 PyTorch model card, 2026. URL https://huggingface.co/google/timesfm-3.0-pytorch. Accessed 2026-09-05.

Alexandre Janon, Thierry Klein, Agnès Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two Sobol index estimators. *ESAIM: Probability and Statistics*, 18: 342–364, 2014. doi: 10.1051/ps/2013040. URL https://arxiv.org/abs/1303.6451.

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. URL https://arxiv.org/abs/2511.11698v3.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317. PMLR, 2025b. URL https://proceedings.mlr.press/v267/liu25be.html.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307, 2008. doi: 10.1186/1471-2105-9-307. URL https://link.springer.com/article/10.1186/1471-2105-9-307.

Qitai Tan, Yiyun Chen, Mo Li, Ruiwen Gu, Yilin Su, and Xiao-Ping (Steven) Zhang. SynTSBench: Rethinking temporal pattern learning in deep learning models for time series. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 161198–161242. Curran Associates, Inc., 2025. doi: 10.52202/085713-4860. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d60361f3d0021c76626b2597957261d6-Paper-Datasets_and_Benchmarks_Track.pdf.

David S. Watson and Marvin N. Wright. Testing conditional independence in supervised learning algorithms. *Machine Learning*, 110(8):2107–2129, 2021. doi: 10.1007/s10994-021-06030-6. URL https://link.springer.com/article/10.1007/s10994-021-06030-6.

James Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Deisenroth. Efficiently sampling functions from Gaussian process posteriors. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10292–10302, 2020. URL https://proceedings.mlr.press/v119/wilson20a.html.

Annan Yu, Danielle C. Maddix, Boran Han, Xiyuan Zhang, Abdul Fatir Ansari, Oleksandr Shchur, Christos Faloutsos, Andrew Gordon Wilson, Michael W. Mahoney, and Yuyang Wang. Understanding the implicit biases of design choices for time series foundation models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 71998–72049, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/753d9584b57ba01a10482f1ea7734a89-Paper-Conference.pdf.

For the mathematical construction, start with Appendix A; for the experimental designs and condition differences, see Appendices B and C.1. Appendices D.1 and E.1 explain residuals and uncertainty. The accompanying *Complete supplementary results* retains the full tables and secondary studies; `RESULTS_INDEX.md` in the anonymous package locates them.

A Conditional sampling and error identities

A.1 Exact conditional replicas

Fix the process parameters, layout and history length. Let $X \sim N(0, \Sigma)$, $\Sigma \succ 0$, and $Y_n \mid X \sim N(S, \sigma_F^2)$, where the true predictive mean $S = \mu_F(X)$ is linear and $\sigma_F^2 > 0$ is constant within the condition. For a full-row-rank constraint matrix C, draw $U \sim N(0, \Sigma)$ independently of (X, Y_n) and set

$$\mathsf{G} = \Sigma C^\top (C\Sigma C^\top)^{-1}, \qquad X^C = \mathsf{G}CX + (I - \mathsf{G}C)U. \tag{10}$$

Since $C\mathsf{G} = I$, $CX^C = CX$. Given $CX = b$, the replica has mean $\mathsf{G}b$ and covariance $\Sigma - \Sigma C^\top (C\Sigma C^\top)^{-1} C\Sigma$, the Gaussian conditional covariance. Independence of the donor makes X^C and X conditionally iid. Integrating the conditional kernel preserves the input marginal. If the constraints determine S, the target law is also unchanged, so (X^C, Y_n) and (X, Y_n) have the same joint distribution. Every fixed pipeline therefore has zero population risk change under this replacement, although its forecasts on individual pairs can differ.

For $Y_t = aY_{t-1} + bY_{t-L} + \varepsilon_t$,

$$S = aX_{n-1} + bX_{n-L}. \tag{11}$$

An S-conditioned replica fixes the true predictive mean; adding the row for the last observation A gives an (A, S)-conditioned replica, with $\mathcal{B} = (A, S)$. Both coefficients are nonzero, so fixing either predictive lagged input and S fixes both inputs. In the layout controls, A denotes the final input value: X_{n-1} in chronological order and X_{n-L} after phase concatenation. An A-conditioned replica X^A fixes only this value and can change the predictive distribution. No model receives the process parameters or true predictive mean.

A.2 Projection and refinement of the state

For a square-integrable deterministic median pipeline f, define $g_S = \mathbb{E}[f \mid S]$, $g_{\mathcal{B}} = \mathbb{E}[f \mid \mathcal{B}]$ and

$$\begin{aligned} V_{\mathcal{B}} &= \mathbb{E}[(f - g_{\mathcal{B}})^2], & V_S &= \mathbb{E}[(f - g_S)^2], \\ D_{\mathcal{B}} &= \mathbb{E}[(g_{\mathcal{B}} - S)^2], & D_S &= \mathbb{E}[(g_S - S)^2], \qquad H = \mathbb{E}[(g_{\mathcal{B}} - g_S)^2]. \end{aligned} \tag{12}$$

The three terms in $f - S = (f - g_{\mathcal{B}}) + (g_{\mathcal{B}} - g_S) + (g_S - S)$ are orthogonal: condition on $\mathcal{B}$ to eliminate the first against either remaining term, and use $\mathbb{E}[g_{\mathcal{B}} - g_S \mid S] = 0$ for the last cross term. Thus

$$\mathcal{E}_{\mathrm{MSE}} = \mathbb{E}[(f - S)^2] = V_{\mathcal{B}} + H + D_S = V_S + D_S = V_{\mathcal{B}} + D_{\mathcal{B}}. \tag{13}$$

Here $V_{\mathcal{B}}$ is remaining history variation, H is conditional mean variation associated with A at fixed S, and D_S is error of the mean forecast given S. Their sum excludes innovation noise. Conditional iid replicas also give

$$V_{\mathcal{B}} = \tfrac{1}{2}\mathbb{E}[(f(X) - f(X^{\mathcal{B}}))^2] = \inf_g \mathbb{E}[(f - g(\mathcal{B}))^2]. \tag{14}$$

These standard conditional-projection identities concern approximation of model outputs, not an unavoidable target-prediction error. For example, $f = S + \lambda(A - \mathbb{E}[A \mid S])$ has $V_{\mathcal{B}} = D_S = 0$ but $D_{\mathcal{B}} = V_S = H = \lambda^2 \mathbb{E}\,\mathrm{Var}(A \mid S)$: the same error is assigned to different components under the two states.

Let $e = f(X) - S$, $e_{\mathcal{B}} = f(X^{\mathcal{B}}) - S$ and $e_S = f(X^S) - S$. The unbiased paired and cross-product estimators are

$$\begin{aligned} \widehat{V}_{\mathcal{B}} &= \mathrm{Avg}[\tfrac{1}{2}(e - e_{\mathcal{B}})^2], & \widehat{V}_S &= \mathrm{Avg}[\tfrac{1}{2}(e - e_S)^2], \\ \widehat{D}_{\mathcal{B}} &= \mathrm{Avg}[e e_{\mathcal{B}}], & \widehat{D}_S &= \mathrm{Avg}[e e_S], \qquad \widehat{H} = \mathrm{Avg}[e(e_{\mathcal{B}} - e_S)]. \end{aligned} \tag{15}$$

The two replicas may share a donor: each must be conditionally independent of its base, but mutual independence of all three outputs is unnecessary. These are pick–freeze conditional-moment estimators (Goda, 2017). The exact sample identity is

$$\widehat{\mathcal{E}}_{\mathcal{B}} = \mathrm{Avg}[(e^2 + e^2_{\mathcal{B}})/2] = \widehat{V}_{\mathcal{B}} + \widehat{H} + \widehat{D}_S. \tag{16}$$

The S-conditioned paired total uses e^2_S instead. The total from original forecasts $\widehat{\mathcal{E}}_0 = \mathrm{Avg}[e^2]$ is common to both state comparisons. All estimate the same population error but need not coincide in sample; neither $\widehat{V}_S - \widehat{V}_{\mathcal{B}} = \widehat{H}$ nor a three-component sum for $\widehat{\mathcal{E}}_0$ is asserted. Signed cross products are retained without clipping. For Sundial, all expectations and projection classes additionally condition on the recorded prediction stream R; these are not averages over fresh prediction seeds.

A.3 Observed-history state for latent sums

The latent extension combines independent unit-variance AR(1) and seasonal-AR components, with seasonal variance share λ. Its covariance is

$$\gamma_h = (1-\lambda)a^h + \lambda \mathbf{1}_{\{L|h\}} b^{\lfloor h/L \rfloor}, \qquad \gamma_0 = 1. \tag{17}$$

For zero-based history indices, set $\Sigma_{ij} = \gamma_{|i-j|}$ and $k_i = \gamma_{n-i}$. Gaussian conditioning gives

$$w = \Sigma^{-1}k, \quad S = w^\top X, \quad \sigma_F^2 = 1 - k^\top w, \quad C = (e_j, w)^\top. \tag{18}$$

Here e_j selects $j = n-1$ for chronology or $j = n-L$ for phase concatenation; resampling occurs in original coordinates before permutation. The sampler uses observed covariances, not realized latent states. The exploratory panel has eight conditions ($a = .55$, $b = \pm.75$, $\lambda = .30, .70$, $L = 11, 23$), six checkpoints, 768 pairs per condition and length 2048. It compares chronological and phase-concatenated layouts, which fix different final values; 16 steps are requested and only the first scored. Complete results remain in the supplementary record.

B Two-lag experiments and normalization

Design and statistical families. The three-arm discovery panel shares the eight conditions, original and (A, S)-conditioned inputs, checkpoints, lengths and 768 tuples per condition with the AR baseline study; it also includes an S-conditioned replica. The confirmation panel crosses $L = 19, 41$, coefficient budgets $c = .55, .85$, local fractions $r = .3, .7$, and both seasonal signs, with $a = cr$, $b = \pm c(1-r)$ and unit marginal variance. The 16 parameter combinations are distinct from discovery within the same Gaussian family. Each condition has 128 tuples per seed across three seeds: 384 tuples, 6,144 in total. All six checkpoints share length-2048 histories and their length-512 suffixes. Requests and scores are one step; confirmation uses independently generated inputs.

The protocol fixed conditions, sample size and six primary endpoints before confirmation inference, without public preregistration or a formal power guarantee. With $\delta = 2048 - 512$, the four component directions are $K_{V,T} = \delta V_T^{\mathrm{Bolt}} - \delta V_T^{\mathrm{Moirai}} > 0$ and $K_{D,T} = \delta D_T^{\mathrm{Moirai}} - \delta D_T^{\mathrm{Bolt}} > 0$, $T \in \{S, \mathcal{B}\}$; the other two require a negative $\delta\mathcal{E}_0$ for original forecasts. Ten thousand stratified paired draws share weights across all models, arms, lengths and variants. The six-endpoint family uses tails $.05/(2 \times 6)$. Only $K_{V,\mathcal{B}}$, $K_{D,\mathcal{B}}$ and Bolt's base-error direction are supported: 3/6, so the criterion requiring all six predictions is not met. The contrasts are algebraically related, not independent replications.

B.1 Checkpoint configuration for the central experiments

The six checkpoint IDs are `amazon/chronos-2`, `amazon/chronos-bolt-base`, `google/timesfm-3.0-pytorch`, `google/timesfm-2.5-200m-pytorch`, `Salesforce/moirai-2.0-R-small` and `thuml/sundial-base-128m`. Model cards document the Bolt and current TimesFM checkpoints (Amazon, 2024; Google, 2025; 2026). Bolt-base has about 205M parameters (Amazon, 2024), versus 11.4M for Moirai-2.0-R-small (Liu et al., 2025a, Table 1); their comparison does not isolate model size from architecture or training. Checkpoint revisions, runtime versions and full wrapper settings are listed in the accompanying *Complete supplementary results* and code.

Table 4: Six prespecified confirmation endpoints on 16 fresh conditions. Values and intervals are in 10^{-3} squared target units.

Endpoint	Direction	Estimate adjusted interval	Supported
Conditioning on S			
Variation contrast $K_{V,S}$	> 0	-2.09 $[-4.95, +0.75]$	No
Mean-error contrast $K_{D,S}$	> 0	-0.90 $[-4.21, +2.54]$	No
Conditioning on (A, S)			
Variation contrast $K_{V,\mathcal{B}}$	> 0	$+12.92$ $[+11.09, +14.71]$	Yes
Mean-error contrast $K_{D,\mathcal{B}}$	> 0	$+13.85$ $[+10.31, +17.43]$	Yes
Original-forecast total error			
Bolt $\delta\widehat{\mathcal{E}}_0$	< 0	-3.75 $[-6.89, -0.61]$	Yes
Moirai $\delta\widehat{\mathcal{E}}_0$	< 0	-2.23 $[-5.73, +1.55]$	No

$\delta = 2048-512$. For either state T, $K_{V,T} = \delta V_T^{\mathrm{Bolt}} - \delta V_T^{\mathrm{Moirai}}$ and $K_{D,T} = \delta D_T^{\mathrm{Moirai}} - \delta D_T^{\mathrm{Bolt}}$. Intervals use bootstrap tails $0.05/(2 \times 6)$ for a nominal 95% family. Complete replication requires all six directed intervals; the contrasts are not independent replications. Original-forecast totals use the same forecasts across states.

Context handling. Each input is a chronological suffix of the requested length. Chronos-2's configured limit is 8192 and Bolt's is 2048. TimesFM 2.5 compiles its maximum context to the next power of two (512 or 2048 here); Moirai sets `context_length=n`; Sundial's wrapper limit is 2880. Chronos, Moirai and Sundial use 16-value patches; TimesFM uses 32-value patches. All central lengths are patch-aligned and retain every observed value without truncation. This was checked in the wrappers and library source; runtime forward hooks were not recorded.

Outputs, randomness and normalization. We score only the first forecast step, using quantiles $\tau = .1, .2, \ldots, .9$ and the median for point analysis. TimesFM 2.5's leading mean channel is discarded; Sundial's quantiles are extracted from 100 samples per tuple with NumPy's linear interpolation. Prediction seed 2026090701 is reset for each generation-seed block, preserving stream slots across arms and lengths without independent prediction-seed replication. The population-normalization variants replace Bolt's and Moirai's internal scaler statistics with mean zero and scale one, including the inverse transformation, while preserving checkpoint weights.

Training context and synthetic data. Configured input limits do not specify the distribution of history lengths seen during training. Chronos-2 was first trained with a maximum context of 2048, then extended to 8192; its synthetic generators include autoregressive models and KernelSynth (Ansari et al., 2025, Sections 3.3 and 4). Moirai 2.0 also includes KernelSynth Gaussian-process series in pretraining (Liu et al., 2025a, Section 4). Thus Gaussian or autoregressive inputs are not automatically outside these models' training distributions. Our controlled processes expose forecast behavior; they do not establish out-of-distribution failure for every checkpoint.

B.2 Component levels and contrasts

The secondary contrast in the last-observation component is $J = \delta H^{\mathrm{Bolt}} - \delta H^{\mathrm{Moirai}} = -.0147473$, with descriptive 95% interval $[-.0164493, -.0130517]$. The signed cross-product estimators give $K_{D,S} - K_{D,\mathcal{B}} = J$ exactly in sample; the analogous variance relation is a population identity. Neither unsupported S-conditioned direction establishes equivalence. All native V_S mean changes are negative, yet Sundial's base-error point estimate increases. Moirai's H increases in all 16 condition point estimates. Table 5 reports all native checkpoints; the supplementary record gives complete levels, condition changes and leave-one-condition-out comparisons.

Table 5: Native checkpoint error changes from 512 to 2048 on the confirmation panel (16 fixed conditions). Entries and descriptive pointwise 95% paired intervals are in 10^{-3} squared target units. H separates last-observation conditional-mean variation from the conditional mean error given S D_S. The common original-forecast total $\widehat{\mathcal{E}}_0$ estimates the same population error as the paired totals, but need not equal their finite-sample component sum. Confirmatory decisions use the separate adjusted endpoint table; no component or model is selected by significance.

Checkpoint	$\delta V_{\mathcal{B}}$	δH	δD_S	$\delta\widehat{\mathcal{E}}_0$
	-8.22	-2.04	-12.92	-21.99
Chronos-2	$[-10.33, -6.14]$	$[-3.65, -0.44]$	$[-15.68, -10.19]$	$[-25.50, -18.40]$
	-3.02	-4.68	$+3.74$	-3.75
Chronos-Bolt	$[-3.80, -2.27]$	$[-5.96, -3.41]$	$[+2.01, +5.51]$	$[-6.05, -1.43]$
	-4.11	-6.09	-17.65	-27.01
TimesFM 3.0	$[-5.77, -2.40]$	$[-7.41, -4.80]$	$[-20.01, -15.28]$	$[-30.14, -23.96]$
	-4.46	-4.89	-4.85	-13.80
TimesFM 2.5	$[-6.15, -2.75]$	$[-6.12, -3.65]$	$[-7.27, -2.45]$	$[-16.97, -10.62]$
	-15.94	$+10.07$	$+2.85$	-2.23
Moirai 2.0	$[-17.10, -14.79]$	$[+8.87, +11.29]$	$[+1.06, +4.70]$	$[-4.92, +0.47]$
	-4.02	$+0.01$	$+7.63$	$+3.47$
Sundial	$[-5.40, -2.66]$	$[-1.78, +1.84]$	$[+4.66, +10.70]$	$[-0.48, +7.56]$

Table 6: Normalization effects at 2048 observations on 16 fresh conditions: native minus population $(0, 1)$. Positive values mean a lower metric under population normalization.

Model	V_S adjusted interval	D_S 95% interval	Base excess MSE 95% interval	Base excess Q9 95% interval
	$+0.97$	-0.85	$+0.09$	$+0.06$
Bolt	$[+0.49, +1.44]$	$[-1.34, -0.34]$	$[-0.62, +0.80]$	$[-0.18, +0.31]$
	-0.56	$+0.34$	-0.25	-0.09
Moirai	$[-0.74, -0.38]$	$[+0.15, +0.52]$	$[-0.51, +0.01]$	$[-0.17, -0.01]$

V_S, D_S and base excess MSE are in 10^{-3} squared target units; Q9 is in 10^{-3} Q9 units. V_S intervals use a separate two-endpoint family (bootstrap tails $0.05/(2 \times 2)$, nominal 95% family coverage). Other intervals are descriptive pointwise 95%. Population statistics use privileged process information.

B.3 Normalization and fitted AR baselines

Bolt's instance-normalization statistics and inverse, and Moirai's packed scaler location and scale, are replaced by population center zero and scale one. This fixed population normalization is applied separately to all three arms at both lengths. The two primary native-minus-population V_S directions at 2048 are expected positive, using tails $.05/(2 \times 2)$. Only Bolt is supported; Moirai moves in the opposite direction. Bolt's D_S increases under the modification, while its base-error and Q9 intervals span zero. Population statistics are privileged information, and these pipeline effects do not isolate a unique internal cause or establish a usable forecasting improvement.

The five fitted AR baselines each estimate their own coefficients and intercept from each input arm. BIC and matched known-lag fits share $t \geq 64$ response rows, giving 448 and 1984 responses at the two lengths. Full-history known-lag fits use $n - L$ responses; a fixed window w uses $\min(n, w) - L$. Only named known-lag references receive L, and no reference receives true coefficients, oracle means or future targets. They provide no Q9 forecast. Fixed-window forecasts are identical across nested lengths when the supplied window is unchanged.

At 2048, the matched BIC and known-lag fits have original-forecast excess error .00096, versus .00095 for the full-history known-lag fit. Restricting the known-lag fit to the last 128 or 512 values gives .02137 or .00403. Dense-reference comparisons appear in Appendix D.5; the supplementary record retains all sparse-reference components and intervals.

Table 7: Component changes across all three panels, 2048 minus 512. Entries and post hoc pointwise 95% paired intervals use 10^{-3} squared target units. Each panel averages its fixed conditions equally. The base total is estimated separately from the paired component sum. Panel comparisons do not isolate the effect of parent structure.

Panel / checkpoint	$\delta V_{\mathcal{B}}$	δH	δD_S	$\delta \mathcal{E}_0$
	$+2.36$	-9.68	$+0.90$	-6.04
Two-lag discovery / Bolt	$[+1.23, +3.47]$	$[-10.93, -8.44]$	$[-0.86, +2.69]$	$[-8.57, -3.51]$
	-17.74	$+10.08$	-1.62	-9.47
Two-lag discovery / Moirai	$[-18.91, -16.59]$	$[+9.01, +11.17]$	$[-3.35, +0.10]$	$[-11.82, -7.16]$
	-3.02	-4.68	$+3.74$	-3.75
Two-lag confirmation / Bolt	$[-3.80, -2.27]$	$[-5.96, -3.41]$	$[+2.01, +5.51]$	$[-6.05, -1.43]$
	-15.94	$+10.07$	$+2.85$	-2.23
Two-lag confirmation / Moirai	$[-17.10, -14.79]$	$[+8.87, +11.29]$	$[+1.06, +4.70]$	$[-4.92, +0.47]$
	-3.26	-5.48	$+2.15$	-6.78
Three-lag / Bolt	$[-4.15, -2.37]$	$[-6.33, -4.65]$	$[+0.97, +3.34]$	$[-8.55, -5.07]$
	-12.03	$+4.53$	-3.03	-10.22
Three-lag / Moirai	$[-12.99, -11.03]$	$[+3.81, +5.24]$	$[-4.27, -1.77]$	$[-11.87, -8.55]$

C TESTING THE RESPONSE TRENDS IN THREE-LAG PROCESSES

C.1 PROSPECTIVE THREE-LAG DESIGN AND CONDITION DIFFERENCES

The design crosses $(L_1, L_2) \in \{(11, 23), (17, 43)\}$, $c \in \{.60, .80\}$, $r \in \{.30, .70\}$ and $s \in \{-1, +1\}$, with $(a, b, d) = (cr, .60sc(1 - r), .40c(1 - r))$. All coefficients are nonzero, and the absolute sum is $c < 1$. The stationary covariance is obtained from the companion Lyapunov equation and scaled to unit marginal variance. The true predictive mean is $S = aA + bP_1 + dP_2$, with independent Gaussian innovation. Fixing (A, S) leaves positive conditional variation in the remaining predictive inputs in every condition.

The three generation seeds are 95201, 95211 and 95221; each supplies 128 independent base/donor tuples per condition. Lengths 128, 256, 512, 1024 and 2048 are nested suffixes at a common target time. Both models receive original, (A, S)-conditioned and S-conditioned inputs at every length. The prospective tests specify $J < 0$, positive Moirai $\delta\alpha$, and negative Bolt $\delta\alpha$, all for 512–2048. The 10000 joint tuple bootstrap draws use seed 95231 and tails $.05/6$ for the three primary intervals. These are related summaries of forecast behavior.

Figure 5 shows all five lengths, and Table 8 reports all 16 condition changes. For the descriptive predictive-mean response, $\delta\beta^{\text{Bolt}} = -.00411$ $[-.01015, .00197]$ and $\delta\beta^{\text{Moirai}} = +.03019$ $[+.02363, +.03661]$. Bolt's decrease in two-lag confirmation is therefore unresolved here. Population-normalization variants and other checkpoints were outside this prospective design.

C.1.1 EXPLORATORY DEPENDENCE ON COEFFICIENT ALLOCATION AND LENGTH

We split the design by the local coefficient share r after examining the results. Each group contains eight conditions, balanced over the other factors. The following intervals are exploratory pointwise 95% intervals from paired tuple resampling within seed and condition. They describe these fixed groups; conditions are not sampled from a wider population.

Bolt's $\delta\alpha$ is $-.0480$ $[-.0533, -.0428]$ at $r = .3$ and $+.0027$ $[-.0112, +.0167]$ at $r = .7$, with a direct group difference of $+.0507$ $[+.0360, +.0655]$. The latter group combines positive and negative condition changes; its small average does not establish equivalence. Moirai's slope increases in both groups: $+.0186$ $[+.0133, +.0239]$ and $+.0563$ $[+.0408, +.0723]$, respectively, with a difference of $+.0377$ $[+.0212, +.0545]$.

Moirai's full length curve is nonmonotone. Its 128–2048 slope change is $+.0002$ $[-.0141, +.0140]$, whereas its 256–2048 change is $+.0330$ $[+.0219, +.0440]$. These additional contrasts do not locate a turning point or replace the prospective 512–2048 comparison.

Table 8: Point-estimate changes from 512 to 2048 in every three-lag condition. H changes use 10^{-3} squared units. Conditions list $L_1, L_2; c, r, s$. Agreement in sign across conditions does not establish significance for any individual condition.

Condition	δH^{Bolt}	δH^{Moirai}	$\delta\alpha^{\text{Bolt}}$	$\delta\alpha^{\text{Moirai}}$
11, 23; .60, .30, -1	-11.070	$+1.714$	-0.038	$+0.011$
11, 23; .60, .30, +1	-9.818	$+2.243$	-0.053	$+0.002$
11, 23; .60, .70, -1	$+1.460$	$+4.477$	$+0.031$	$+0.075$
11, 23; .60, .70, +1	-1.638	$+0.636$	-0.025	$+0.039$
11, 23; .80, .30, -1	-9.644	$+10.454$	-0.038	$+0.027$
11, 23; .80, .30, +1	-15.514	$+6.525$	-0.057	$+0.018$
11, 23; .80, .70, -1	-0.678	$+2.720$	$+0.003$	$+0.067$
11, 23; .80, .70, +1	-3.219	$+4.022$	-0.044	$+0.049$
17, 43; .60, .30, -1	-7.635	$+2.593$	-0.043	$+0.006$
17, 43; .60, .30, +1	-11.193	$+0.906$	-0.063	$+0.013$
17, 43; .60, .70, -1	-1.592	$+5.257$	-0.029	$+0.109$
17, 43; .60, .70, +1	$+1.481$	$+1.946$	$+0.031$	$+0.030$
17, 43; .80, .30, -1	-9.272	$+14.002$	-0.039	$+0.035$
17, 43; .80, .30, +1	-11.426	$+8.253$	-0.054	$+0.037$
17, 43; .80, .70, -1	-2.533	$+4.634$	-0.019	$+0.052$
17, 43; .80, .70, +1	$+4.653$	$+2.043$	$+0.073$	$+0.030$

Comparing responses in output units. The standardized response $t_A = \alpha\sqrt{q}$ measures output change over one residual standard deviation of A, computed within each condition before group averaging. Bolt's group difference remains positive: $+.0399$ $[+.0340, +.0458]$. Moirai's changes are $+.0150$ $[+.0107, +.0193]$ at $r = .3$ and $+.0166$ $[+.0120, +.0213]$ at $r = .7$; their difference, $+.0016$ $[-.0046, +.0081]$, is unresolved. Thus the larger raw slope increase at $r = .7$ does not establish a larger increase in output units. Changing r changes both the covariance and the model response, so these comparisons do not isolate either source of the change in H.

Responses along other predictive lagged inputs. For each predictive lagged input $P_j \in \{A, P_1, P_2\}$, define $Z_j = P_j - \text{Cov}(P_j, S)S/v$, $q_j = \text{Var}(Z_j)$, and $\alpha_j = \mathbb{E}[(f - S)Z_j]/q_j$. The full supplementary results report $\text{sign}(\theta_j)\alpha_j\sqrt{q_j}$, where $(\theta_A, \theta_1, \theta_2) = (a, b, d)$; orienting by the coefficient sign prevents cancellation across positive and negative b conditions.

For the combined nonlocal contribution $T = bP_1 + dP_2 = S - aA$, $a \neq 0$ gives $\sigma(A, S) = \sigma(T, S)$ and $T - \mathbb{E}[T \mid S] = -aZ_A$. Fixing (A, S) therefore fixes T while leaving one direction within (P_1, P_2) free. The residual directions satisfy $aZ_A + bZ_1 + dZ_2 = 0$: their response slopes are overlapping projections, not separate causal input weights. Variation conditional on (P_1, S) or (P_2, S) requires its own replicas and cannot be recovered from the (A, S) arm.

D Affine diagnostics and fitted AR baselines

D.1 Decomposing the residual after the two linear responses

Within each Gaussian process, $\mathbb{E}[Z_A \mid S] = 0$ and

$$f - \eta - \beta S - \alpha Z_A = (f - g_{\mathcal{B}}) + (g_{\mathcal{B}} - g_S - \alpha Z_A) + (g_S - \eta - \beta S).$$

The first term is orthogonal to every $\mathcal{B}$-measurable function; the second has conditional mean zero given S and is orthogonal to the third. Their squared norms are $V_{\mathcal{B}}$, $H - H^{\text{lin}}$ and D_S^{rem}, proving Eq. (9). This requires conditional centering of Z_A, not merely zero covariance.

Equal summaries, different conditional components. Let $U = S/\sqrt{v}$, $W = Z_A/\sqrt{q}$ and a standardized history residual T be mutually independent standard Gaussians. Adding λT, $\lambda(U^2 -$

Table 9: Affine residuals and their components at confirmation length 2048. Values and pointwise 95% intervals use 10^{-3} squared units. R_{CV} includes the error of estimating the affine coefficients; N is the nonlinear conditional-mean remainder. All nine predictors are retained. Pop. denotes population normalization.

Predictor	$\widehat{R}_{\mathrm{aff}}$	$\widehat{R}_{\mathrm{CV}}$	$V_{\mathcal{B}}$	N
Chronos 2	34.93 [31.80, 37.95]	35.21 [33.99, 37.30]	35.42 [33.78, 37.11]	−0.07 [−3.70, 3.30]
Chronos-Bolt	17.77 [10.82, 24.27]	19.99 [19.25, 21.26]	18.88 [18.13, 19.66]	0.24 [−6.94, 6.96]
TimesFM 3.0	30.76 [28.02, 33.41]	31.09 [29.97, 33.05]	31.29 [29.84, 32.75]	0.23 [−2.97, 3.25]
TimesFM 2.5	31.66 [27.68, 35.39]	32.93 [31.80, 34.92]	33.79 [32.34, 35.30]	−0.94 [−5.37, 3.26]
Moirai 2.0	16.53 [10.61, 21.97]	17.63 [17.03, 18.75]	17.44 [16.69, 18.21]	−1.39 [−7.85, 4.53]
Bolt (pop.)	16.91 [9.92, 23.45]	19.17 [18.48, 20.40]	17.96 [17.23, 18.71]	0.23 [−6.99, 7.03]
Moirai (pop.)	16.86 [10.94, 22.30]	17.96 [17.34, 19.09]	17.74 [16.97, 18.52]	−1.33 [−7.82, 4.60]
Dense OLS	22.33 [21.44, 23.24]	22.59 [21.92, 23.82]	22.01 [21.20, 22.86]	0.12 [−0.50, 0.72]
Dense ridge	19.44 [18.70, 20.20]	19.68 [19.14, 20.73]	19.16 [18.47, 19.87]	0.17 [−0.38, 0.71]

$1)/\sqrt{2}$ or λUW to the same affine predictor preserves its affine coefficients and adds residual energy λ^2. The three allocations $(V_{\mathcal{B}}, H - H^{\mathrm{lin}}, D_S^{\mathrm{rem}})$ are $(\lambda^2, 0, 0)$, $(0, 0, \lambda^2)$ and $(0, \lambda^2, 0)$. These are analytic examples, not measured checkpoint effects.

Estimating the components. Using only the original forecasts, we estimate the residual energy by $\widehat{R}_{\mathrm{aff}} = \widehat{\mathcal{E}}_0 - \widetilde{H}^{\mathrm{lin}} - \widetilde{D}_S^{\mathrm{affine}}$, where squared moments use distinct-seed products (Appendix D.3). With $\widehat{N} = (\widehat{H} - \widetilde{H}^{\mathrm{lin}}) + (\widehat{D}_S - \widetilde{D}_S^{\mathrm{affine}})$, the exact sample identity is

$$\widehat{R}_{\mathrm{aff}} + (\widehat{\mathcal{E}}_{\mathcal{B}} - \widehat{\mathcal{E}}_0) = \widehat{V}_{\mathcal{B}} + \widehat{N}.$$

The correction accounts for two finite-sample estimates of the same total; it is not additional nonlinear energy. Signed estimates are retained.

Held-out affine regression. Within each process, predictor and length, fit $e = f - S$ to $\phi = (1, S, Z_A)^\top$ on two generation seeds and evaluate the third. Average the squared held-out residuals over the three rotations, then conditions equally. This $\widehat{R}_{\mathrm{CV}}$ includes coefficient-estimation error and differs from the population-projection target of $\widehat{R}_{\mathrm{aff}}$. Each of 10000 draws resamples complete tuples within condition and seed, refits training coefficients, and shares weights across held-out losses, replica components, lengths and predictors.

Table 9 reports confirmation at 2048; the accompanying *Complete supplementary results* retains both panels, both lengths and paired changes. These post hoc comparisons use pointwise 95% intervals for fixed conditions. Sundial is excluded from these deterministic projections. Native-checkpoint intervals for N span zero throughout the full comparison. Dense ridge is an exception at 512 in confirmation: $\widehat{N} = .00212$ [.00026, .00392], with change $-.00195$ [$-.00373, -.00015$], primarily in D_S^{rem}. Chronos-2 and TimesFM 3.0 have positive intervals for $H - H^{\mathrm{lin}}$ at 2048, but their nonlinear sums are not established as positive.

A scalar-calibrated OLS reference. To ask whether lower history variation alone explains the comparison, we use the saved dense-OLS forecasts and consider $\widetilde{f}_\lambda = \lambda f_{\mathrm{OLS}}$, with $\lambda = \sqrt{V_{\mathcal{B}}^{\mathrm{checkpoint}}/V_{\mathcal{B}}^{\mathrm{OLS}}}$ at each bootstrap draw. This is a post hoc descriptive reference: it scales outputs around zero, uses no new calibration set, and is not a proposed forecasting method. Table 10 shows the matched comparisons. The native checkpoints retain much larger excess error at the same aggregate $V_{\mathcal{B}}$; this comparison does not identify an internal mechanism or rule out other shrinkage families.

Share of the residual due to nonlinear conditional means. For each fixed panel, predictor and length, define

$$\rho = \frac{\overline{N}}{\overline{V_{\mathcal{B}} + N}}, \qquad N = \mathbb{E}[(g_{\mathcal{B}} - \eta - \beta S - \alpha Z_A)^2].$$

Table 10: A descriptive scalar-calibrated OLS reference at confirmation length 2048. Energy columns use 10^{-3} squared target units. We scale dense OLS by λ to match each checkpoint's aggregate $V_{\mathcal{B}}$, then compare excess error. The selected λ is recomputed in every paired bootstrap draw; intervals for the error gap are pointwise and exploratory.

Model	λ	Matched $V_{\mathcal{B}}$	Native $\mathcal{E}_0$	Scaled OLS $\mathcal{E}_0$	Native − scaled
Bolt	0.926	18.88	154.29	21.02	133.27 [126.40, 140.08]
Moirai	0.890	17.44	133.13	21.80	111.32 [105.19, 117.47]

Table 11: Nonlinear mean residual share $\rho = \overline{N}/(\overline{V}_{\mathcal{B}} + \overline{N})$, in percent. D and C denote discovery and confirmation. Brackets give exploratory nominal pointwise 95% intervals for every native checkpoint and both lengths; negative endpoints are retained. U_5 is the one-sided nominal 95% upper bound at C2048, Bonferroni-adjusted over the five checkpoints using the 0.99 draw quantile. This family excludes other panels and lengths. The ratio uses condition-averaged energies, with the resampling-based affine residual in the denominator; it is not a fraction of total forecast error or an average of per-condition ratios. These post hoc bounds are not a prospective equivalence test.

Predictor	D512	D2048	C512	C2048	C2048 U_5
Chronos 2	$[-11.9, 8.4]$	$[-7.4, 8.7]$	$[-8.5, 10.5]$	$[-11.4, 8.6]$	10.0
Chronos-Bolt	$[-21.3, 15.9]$	$[-18.7, 14.0]$	$[-38.7, 26.8]$	$[-58.0, 27.1]$	29.7
TimesFM 3.0	$[-8.1, 9.2]$	$[-8.7, 7.0]$	$[-8.5, 13.6]$	$[-10.4, 9.6]$	11.0
TimesFM 2.5	$[-7.9, 9.3]$	$[-9.4, 9.9]$	$[-14.0, 10.7]$	$[-18.7, 8.9]$	10.7
Moirai 2.0	$[-16.6, 10.6]$	$[-39.1, 17.9]$	$[-23.3, 12.8]$	$[-81.9, 20.9]$	24.4

Bars average energies over conditions before division; this is not an average of within-condition ratios or a fraction of total forecast error. Each paired bootstrap draw divides $\widehat{N}$ by $\widehat{V}_{\mathcal{B}}+\widehat{N}$, including the sampling correction above. All draw denominators are positive; signed estimates and intervals are retained without clipping. Table 11 gives both panels and lengths. Its confirmation-2048 upper bounds use the .99 quantile, a one-sided Bonferroni adjustment at nominal 95% across these five checkpoints only. The analysis is exploratory; ratio coverage has not been separately calibrated and the adjustment does not cover selection across panels, lengths or other analyses. A small share constrains the nonlinear conditional mean $g_{\mathcal{B}}$; it neither makes f affine nor establishes equivalence to a fitted linear forecaster.

D.2 Paired residuals: lower variation with a different target

In the saved forecasts, fitting the linear response first reduces sampling variation but changes the target. Fit $\widehat{\ell}_{-k}$ on two generation-seed folds and average $(f(X) - \widehat{\ell}_{-k}(\mathcal{B}))(f(X^{\mathcal{B}}) - \widehat{\ell}_{-k}(\mathcal{B}))$ on the excluded fold. Conditional independence of the two forecasts gives

$$Q_k = \mathbb{E}[(g_{\mathcal{B}} - \widehat{\ell}_{-k})^2 \mid \widehat{\ell}_{-k}] = N + \|\widehat{\ell}_{-k} - \ell_*\|_2^2 \geq N, \qquad Q = \tfrac{1}{3}\sum_{k=1}^{3} Q_k,$$

where ℓ_* is the population affine projection. Orthogonality removes the cross term, with the squared norm taken over $\mathcal{B}$. The signed estimate $\widehat{Q}$ is not a guaranteed upper bound for N, and a positive interval can reflect fitting error. Training sets overlap across rotations, so all fits are recomputed together in each of the 10000 complete-tuple bootstrap draws.

At confirmation length 2048, bootstrap standard deviations of $\widehat{Q}$ are 8.4–33.4% of those of $\widehat{N}$ for the five native checkpoints, but 99.4% and 94.6% for OLS and ridge. These are saved-forecast comparisons; complete estimates appear in the accompanying *Complete supplementary results*.

Calibration exposes undercoverage. With independent standard Gaussian S, Z, T, T', the simulated predictor is $f = .1 + .6S + .3Z + \lambda(S^2 - 1)/\sqrt{2} + \kappa SZ + \tau T$; its replica replaces T by T'.

Table 12: Lower estimation error does not ensure nominal coverage. Coverage of N by nominal one-sided 95% percentile endpoints, in percent, with Wilson 95% Monte Carlo intervals. Each case uses 500 independent single-condition panels and 999 jointly refitted bootstrap draws per panel. The last column is $100\,\mathrm{RMSE}(\widehat{Q}, N)/\mathrm{RMSE}(\widehat{N}, N)$.

Analytic case	$\widehat{Q}$ coverage	$\widehat{N}$ coverage	RMSE ratio (%)
Linear mean + history variation	98.6 [97.1, 99.3]	96.2 [94.1, 97.6]	9.4
Nonlinear mean only	**90.4** [87.5, 92.7]	95.6 [93.4, 97.1]	33.8
Nonlinear mean + history variation	94.0 [91.6, 95.8]	95.0 [92.7, 96.6]	35.1
Small nonlinear mean + larger variation	99.0 [97.7, 99.6]	95.4 [93.2, 96.9]	20.3

Thus $N = \lambda^2 + \kappa^2$ and $V_{\mathcal{B}} = \tau^2$. In table order, (λ, κ, τ) equals $(0, 0, .2)$, $(.15, .1, 0)$, $(.15, .1, .2)$ and $(.04, .03, .3)$. Each case uses 500 independent panels of 3×128 tuples and 999 jointly refitted bootstrap draws per panel. **With a nonlinear mean but no history variation, nominal 95% one-sided coverage of N is only 90.4%**, versus 95.6% for the original estimator (Table 12). Lower RMSE therefore does not ensure valid intervals. These single-condition simulations do not calibrate the 16-condition TSFM averages. This exploratory comparison replaces neither ρ and its bounds nor any confirmation result.

D.3 Last-observation response and variation across conditions

All analyses here are post hoc, with pointwise 95% paired percentile intervals conditional on the fixed processes and checkpoints.

Projection and interpretation. For process θ, let $Z_\theta = A - \mathbb{E}[A \mid S, \theta]$, $q_\theta = \mathbb{E}[Z_\theta^2] > 0$ and $W = g_{\mathcal{B}} - g_S$. Since Z_θ is $\mathcal{B}$-measurable and conditionally centered, $m_\theta := \mathbb{E}[(f - S)Z_\theta] = \mathbb{E}[WZ_\theta]$ and $\alpha_\theta = m_\theta/q_\theta$. Orthogonal projection gives

$$H_{\mathrm{lin},\theta} = m_\theta^2/q_\theta \le H_\theta, \qquad H_\theta - H_{\mathrm{lin},\theta} = \mathbb{E}[(W - \alpha_\theta Z_\theta)^2]. \tag{19}$$

For $S = aA + bP$, the known Gaussian covariance gives

$$\gamma_{AS,\theta} = a\gamma_0 + b\gamma_{L-1}, \qquad v_\theta = (a^2 + b^2)\gamma_0 + 2ab\gamma_{L-1},$$
$$Z_\theta = A - (\gamma_{AS,\theta}/v_\theta)S, \qquad q_\theta = \gamma_0 - \gamma_{AS,\theta}^2/v_\theta.$$

Here $\gamma_{AS,\theta} = \mathrm{Cov}(A, S \mid \theta)$. We compute residuals and squared projections within conditions before averaging equally. For nonzero a, b, $\sigma(A, S) = \sigma(P, S)$, $Z_P = -(a/b)Z_A$ and $\alpha_P = -(b/a)\alpha_A$, so the signed coefficients change coordinates while their linear energies agree. For example, $f = bP$ has $\alpha_A = -a$ without explicitly using A. Conversely, $f = S + \kappa S Z_\theta$ has $\alpha = 0$ but $H > 0$ for independent centered Gaussian S, Z_θ of positive variance. Thus α neither identifies a direct input weight nor captures all of H.

Independent-fold energy estimator. Let $\widehat{m}_k$ be the mean of $(f - S)Z_\theta$ in generation-seed fold k. The all-fold mean divided by q_θ estimates α_θ, and

$$\widetilde{H}_{\mathrm{lin},\theta} = \frac{\sum_{k \ne l} \widehat{m}_k \widehat{m}_l}{3 \cdot 2\, q_\theta} \tag{20}$$

is unbiased because distinct folds are independent. Squaring one sample moment instead adds its variance divided by q_θ. The shared 10000 within-condition/seed tuple draws recompute all moments and cross products. We do not force estimated energies to be nonnegative: discovery Moirai at $n = 128$ has $\widehat{H} - \widetilde{H}^{\mathrm{lin}} = -.00385$ $[-.00711, -.00071]$, an unconstrained estimate rather than a negative population energy. Ratios $\delta H_{\mathrm{lin}}/\delta H$ are not explained fractions. Calibration appears in Appendix E.1.

Coefficient budget and normalization. The response contrasts vary across coefficient budgets; the accompanying *Complete supplementary results* gives the direct group comparisons. Population normalization changes J from $-.01475$ to $-.01307$, with paired difference $+.00168$ $[.00113, .00222]$. The 88.6% magnitude retention is a descriptive point ratio, not a causal mechanism fraction.

D.4 Predictive-mean response and equivalent lag coefficients

Within each process, $1, S, Z_A$ are orthogonal in L^2 and $\mathbb{E}[Z_A \mid S] = 0$. The affine projection of $g_S - S$ is $\eta + (\beta - 1)S$, giving

$$D_S^{\mathrm{rem}} = \mathbb{E}[(g_S - S - \eta - (\beta - 1)S)^2], \qquad D_S = \eta^2 + (\beta - 1)^2 v + D_S^{\mathrm{rem}}.$$

This is orthogonal to the last-observation projection in Eq. (19). We estimate η by $\mathrm{Avg}(f - S)$, using the known zero mean of S as a control variate. Distinct-fold products of the fold means of e, eS and eZ_A, with divisors $1, v, q$, estimate η^2, $(\beta - 1)^2 v$ and $\alpha^2 q$. Signed energy estimates are retained.

Substituting $S = aA + bP$ gives equivalent output-projection coordinates

$$\eta + \beta S + \alpha Z_A = \eta + uA + w_P P, \qquad u = \alpha + a(\beta - \alpha\gamma_{AS}/v), \quad w_P = b(\beta - \alpha\gamma_{AS}/v).$$

Here $\gamma_{AS} = \mathrm{Cov}(A, S)$; these are output projections, not internal network parameters.

In confirmation, Bolt's and Moirai's D_S increases mainly accompany increases in its affine part: .00426 $[.00200, .00659]$ and .00365 $[.00135, .00605]$, respectively. Remaining-change intervals span zero; this does not establish absence. Moirai's mean β and mean squared slope error both rise because conditions differ: the square of the mean slope error need not equal the mean squared error. Complete response and lag-coefficient tables for both panels appear in the accompanying *Complete supplementary results*.

D.5 Fitted dense AR baselines

Dense OLS regresses each history's Y_t on an intercept and $Y_{t-1}, \ldots, Y_{t-64}$ for $t \geq 64$. Ridge adds $T\tau\|w\|^2$ to residual sum of squares (RSS), with $T = n - 64$ and unpenalized intercept. It selects $\tau \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ per history by minimizing $\mathrm{RSS}/(T - \mathrm{df})^2$, where $\mathrm{df} = 1 + \sum_j \lambda_j/(\lambda_j + T\tau)$ and λ_j are centered-design Gram eigenvalues. The grid precedes fitting; selection uses neither the target nor true period. Dense and sparse candidates use the same response rows. At 128 the unregularized dense fit lacks residual degrees of freedom, so these comparisons use 512 and 2048.

All three forecast arms are fitted separately, with intervals from the central bootstrap weights. Both dense model classes contain the true process, an inductive advantage over a general predictor. Table 13 reports confirmation; the discovery comparison appears in the accompanying *Complete supplementary results*. The error ratio is $\sum_\theta w_\theta \mathcal{E}_{0,\theta} / \sum_\theta w_\theta \mathrm{Var}(S_\theta)$, with equal fixed-condition weights; confirmation's denominator is .34495. It measures error relative to oracle-predictable variance, not information used. Sparse AR tests form a separate prespecified family (Appendix F).

Excess error and total forecast error. Total MSE adds innovation noise: $\mathrm{MSE} = \mathcal{E}_0 + \sigma^2$. The mean innovation variance in confirmation is 0.65505. At 2048, the resulting total-MSE point estimates are 0.67744 for OLS, 0.80934 for Bolt and 0.78818 for Moirai. Bolt and Moirai therefore exceed OLS by 19.5% and 16.3% in total MSE, respectively. The six-to-sevenfold comparison concerns excess error, the part above the irreducible innovation variance. These are descriptive point comparisons, without new hypothesis tests.

Baseline responses and remaining history variation. The same affine projection, independent-fold energy estimator and shared-tuple bootstrap apply to the dense AR forecasts. At confirmation length 2048, OLS and ridge give $(\beta, \alpha) = (1.002\,[.994, 1.010], -.007\,[-.014, .001])$ and $(.947\,[.940, .955], -.010\,[-.017, -.002])$, respectively. These are post hoc pointwise 95% intervals. Near-ideal mean coefficients coexist with nonzero remaining history variation, and intervals spanning ideal values do not establish equivalence.

Table 13: The fitted reference changes the comparison. Confirmation at 2048, with post hoc pointwise 95% intervals. History variation $V_{\mathcal{B}}$ uses 10^{-3} squared units; retention is $100V_{\mathcal{B},2048}/V_{\mathcal{B},512}$. The last column divides mean excess MSE by mean predictable variance.

Model / reference	$V_{\mathcal{B},2048}$	Retention (%)	$\mathcal{E}_0/\operatorname{Var}(S)$ (%)
Sparse BIC	0.98 [0.93, 1.03]	16.4 [15.4, 17.6]	0.3 [0.3, 0.3]
Dense AR(64)	22.01 [21.20, 22.86]	20.4 [19.5, 21.4]	6.5 [6.2, 6.8]
Dense ridge (GCV)	19.16 [18.47, 19.87]	38.8 [37.0, 40.6]	5.9 [5.7, 6.1]
Chronos-2	35.42 [33.78, 37.11]	81.2 [77.0, 85.5]	26.6 [25.4, 27.7]
Chronos-Bolt	18.88 [18.13, 19.66]	86.2 [83.0, 89.4]	44.7 [42.7, 46.8]
TimesFM 3.0	31.29 [29.84, 32.75]	88.4 [84.1, 93.0]	24.0 [23.1, 25.1]
TimesFM 2.5	33.79 [32.34, 35.30]	88.3 [84.3, 92.6]	30.2 [28.9, 31.5]
Moirai 2.0	17.44 [16.69, 18.21]	52.2 [50.2, 54.4]	38.6 [36.8, 40.4]
Sundial	22.93 [21.98, 23.91]	85.1 [80.6, 89.8]	43.6 [41.6, 45.7]

E FINITE-SAMPLE CALIBRATION

E.1 PROJECTION RESIDUALS WITH KNOWN TRUTH

A negative residual estimate is possible even when its population value is nonnegative. To check the intervals, we repeat the full analysis on analytic predictors with the same condition variances and sample sizes as discovery (eight conditions, 3×256 tuples each) and confirmation (16 conditions, 3×128). Each design uses 1000 independent panels and 10000 complete-tuple bootstrap draws per panel, recomputing fold products and their joint covariance.

Within a condition, $S \sim N(0, v)$ and $Z_A \sim N(0, q)$ are independent. The linear case is $f = S + .3Z_A$, with both nonlinear residuals zero. The nonlinear case is

$$f - S = .08 - .35S + .3Z_A + .12\frac{S}{\sqrt{v}}\frac{Z_A}{\sqrt{q}} + .12\frac{S^2/v - 1}{\sqrt{2}}.$$

Here $H - H^{\mathrm{lin}} = D_S^{\mathrm{rem}} = .0144$, $H^{\mathrm{lin}} = .09q$ and $D_S^{\mathrm{affine}} = .08^2 + .35^2 v$. The (A, S)-conditioned replica leaves the predictor unchanged; the S-conditioned replica replaces Z_A independently. Gaussian quadrature independently checks these truths.

Table 14 gives residual coverage and Monte Carlo intervals. At the zero boundary, $H - H^{\mathrm{lin}}$ intervals lie wholly below zero in 2.0% and 2.7% of panels. This illustrates the possibility of a negative interval, such as discovery Moirai at length 128, without clipping or reinterpreting it. Coverage of D_S^{affine} at its zero boundary is conservative (99.9% and 99.8%). Joint resampling preserves the substantial covariance between H and H^{lin}; subtracting separately computed interval endpoints would not. The full bias, covariance and tail-miss results are in the supplementary record.

E.2 HISTORY VARIATION AND JOINT RESPONSE TESTS

A second simulation allows $V_{\mathcal{B}} > 0$. Let R be the standardized Gaussian residual of P_1 given (A, S) in each of the 16 three-lag conditions. It is independent of (S, Z_A), so

$$f_n = \beta_n S + \alpha_n Z_A + \lambda_n R \quad \Longrightarrow \quad (V_{\mathcal{B},n}, H_n, D_{S,n}) = (\lambda_n^2, \alpha_n^2 q, (\beta_n - 1)^2 v).$$

The nonlinear case adds $.04 + .12(S/\sqrt{v})(Z_A/\sqrt{q}) + .10(S^2/v - 1)/\sqrt{2}$, adding .0144 to H and .0116 to D_S. Two predictors use $(\beta_{512}, \beta_{2048}) = (.65, .72), (.55, .65)$ and $(\lambda_{512}, \lambda_{2048}) = (.20, .15), (.24, .17)$. The signal cases use $\alpha = (.30, .27), (.24, .28)$; the global null uses $(.30, .30), (.24, .24)$, and the partial boundary uses $(.30, .30), (.24, .28)$. Thus null response tests still allow changes in other components.

Replicas share a donor, and histories at 512 and 2048 share their suffix. Each of 1000 independent panels uses 3×128 tuples per condition and 10000 shared-tuple bootstrap draws. The three response intervals use tails $.05/6$. Table 15 separates simultaneous coverage from the chance that all three directed tests pass: the latter measures power for signals and erroneous complete confirmation for

Table 14: Calibration on 1000 independent simulated panels per design, with 10000 shared-tuple bootstrap draws per panel. H and H_{lin} bias use 10^{-4} units; residual coverage is in percent with Wilson 95% Monte Carlo intervals. The linear case has both residual truths zero; the nonlinear case has both equal to .0144.

Design / predictor	Bias H	Bias H_{lin}	$H - H_{\mathrm{lin}}$ coverage	$D_{S,\mathrm{rem}}$ coverage
Discovery / Linear	$+0.04$	-0.37	94.7 [93.1, 95.9]	94.0 [92.4, 95.3]
Discovery / Nonlinear	-0.06	-0.22	93.8 [92.1, 95.1]	95.3 [93.8, 96.4]
Confirmation / Linear	$+0.10$	$+0.13$	94.4 [92.8, 95.7]	95.4 [93.9, 96.5]
Confirmation / Nonlinear	$+0.15$	$+0.40$	95.1 [93.6, 96.3]	95.5 [94.0, 96.6]

Table 15: Calibration with nonzero remaining history variation. All cases use the 16 three-lag conditions, 3×128 tuples per condition, shared donor replicas, and nested lengths 512 and 2048. Each of 1000 independent panels uses 10000 paired bootstrap draws. Endpoint intervals use tails $.05/6$ (nominal 98.33% marginal, 95% simultaneous). Entries are percentages; brackets give 95% Wilson intervals for Monte Carlo uncertainty. Passing all three directions measures power in signal cases and erroneous complete confirmation in null cases. This finite collection does not establish uniform coverage.

Analytic configuration	Adjusted marginal	Simultaneous coverage	Three directions pass
Linear signal	98.0–98.7	96.0 [94.6, 97.0]	100.0 [99.6, 100.0]
Nonlinear signal	97.7–98.7	95.6 [94.1, 96.7]	100.0 [99.6, 100.0]
Global endpoint null	98.5–99.4	98.0 [96.9, 98.7]	0.0 [0.0, 0.4]
One endpoint at zero	98.0–98.7	96.3 [94.9, 97.3]	0.5 [0.2, 1.2]

the nulls. Ordinary pointwise coverage across 144 level/change statistics is 93.5–95.7%; adjusted simultaneous coverage is 95.6–98.0%. Complete confirmation occurs in 0/1000 global-null and 5/1000 partial-boundary panels.

These finite analytic cases do not establish uniform coverage for arbitrary checkpoint outputs or for the residual-share ratios. The exploratory paired residual estimator has its own calibration, including 90.4% coverage for a nominal 95% upper bound (Appendix D.2).

F FITTED AR COMPARISONS WITH NESTED HISTORIES

The eight discovery conditions cross $L = 13, 29$, both signs of b, and $(a, |b|) = (.25, .45), (.45, .25)$ at unit marginal variance. Each has three seeds of 256 independent base/donor pairs. Length-128 and length-512 inputs are suffixes of the same length-2048 histories, with both predictive lagged inputs fixed by Eq. (10). All scores concern the next step.

The sparse baseline fits an intercept and selects AR(1) or AR$(1, L)$, $L = 2, \ldots, 64$, by $\mathrm{BIC} = (n - 64)\log(\mathrm{RSS}/(n - 64)) + k\log(n - 64)$, with k fitted coefficients and the same response rows $t \geq 64$ for every candidate. Each input arm and length is fitted separately, without true coefficients, predictive means or future targets. Its candidate family contains the true process, unlike a general forecasting model. Known-period and zero-intercept fits are additional references in the accompanying *Complete supplementary results*.

The two prespecified contrasts for each of six checkpoints are

$$G_{\mathrm{len}} = V_{2048} - V_{2048}^{\mathrm{BIC}}, \qquad C_{\mathrm{len}} = \log(V_{2048}/V_{128}) - \log(V_{2048}^{\mathrm{BIC}}/V_{128}^{\mathrm{BIC}}). \tag{21}$$

Here $V = V_{\mathcal{B}}$. The 10000 paired bootstrap draws retain shared tuples across models, lengths and references, with equal condition/seed weights. All twelve intervals, adjusted together using tails $.05/24$, are positive: the checkpoints have greater remaining history variation at 2048 and retain a larger fraction of their length-128 variation than BIC. At 2048, BIC selects the true period in every

input; checkpoint variation is 16.69–35.14 times its value. These finite-history comparisons do not establish asymptotic floors or a unique learning mechanism. Condition effects also differ: variation rises from 512 to 2048 in 10/48 checkpoint–condition point estimates, and Bolt's pooled increase reverses on omitting one condition. All conditions remain in the estimates. The complete supplementary record retains all contrast intervals and component, condition and reference comparisons.

G Q9 RISK AND PAIRED AVERAGING

For $\mathcal{Q} = \{.1, \ldots, .9\}$, the unnormalized pinball score is

$$S_{\mathrm{q9}}(y, q) = \frac{2}{9} \sum_{\tau \in \mathcal{Q}} \rho_\tau(y - q_\tau), \qquad \rho_\tau(u) = u(\tau - \mathbf{1}\{u < 0\}). \tag{22}$$

Summed Q9 is the numerator of WQL (Ansari et al., 2024, Appendix D). For $Y \mid S \sim N(S, \sigma_F^2)$ and $z = (q - S)/\sigma_F$,

$$\ell_\tau(q) = \mathbb{E}[\rho_\tau(Y - q) \mid S] = \sigma_F \phi(z) + (q - S)(\Phi(z) - \tau). \tag{23}$$

Thus $\mathcal{L}_S(q) = (2/9) \sum_\tau \ell_\tau(q_\tau)$ integrates future-innovation noise analytically; its optimum is $q_\tau^* = S + \sigma_F \Phi^{-1}(\tau)$. For conditional replicas q, q', with $\bar{q} = (q + q')/2$, convexity gives

$$0 \le d := \tfrac{1}{2}\{\mathcal{L}_S(q) + \mathcal{L}_S(q')\} - \mathcal{L}_S(\bar{q}) \le \tfrac{1}{2}\{\mathcal{L}_S(q) + \mathcal{L}_S(q')\} - \mathcal{L}_S(q^*). \tag{24}$$

This improvement is relative to mean arm risk; the oracle-dependent replicas do not provide a deployable ensemble. In the eight-condition nested-history panel at 2048, mean gains across six checkpoints range from .00316 to .00679 Q9 units (6.3–21.1% of excess risk); Chronos-2's .00679 is only 1.24% of total mean-arm risk. These post hoc results, their paired uncertainty, and all quantile-spacing and common-shift analyses remain in the complete supplementary record.

H SEPARATE PHASE-PERMUTATION STUDIES AND THEIR LIMITS

These studies ask whether changing the presentation of a history improves forecasts. Phase concatenation groups indices by residue modulo the known period L, preserving chronology within each group and placing the target phase last. It is an invertible permutation: it retains all observations and the optimal achievable risk, but a fixed checkpoint can change its prediction. It is a different intervention from the main paper's conditional resampling. All positive and negative results, designs and controls are retained together in the complete supplementary record.

In the seasonal-AR experiment, permutation lowers first-step Q9 in all 24 long-history cells (six checkpoints, two periods and two coefficient signs). Fitted AR(1) also improves, so this is not specific to foundation models. Requesting one step instead of 16 changes the measured benefit by at most 1.94×10^{-9}; this first-step check does not establish multi-step validity.

The real-background test uses three models and seven previously screened GIFT-Eval sources (Aksu et al., 2024), with 64 series each and native or seasonal-AR-injected targets. All six chronology-minus-permutation means and source-bootstrap intervals are negative. No model passes both injected contrasts, so the rule requiring two of three models stops expansion. Selection and the unknown real-process distribution prevent interpreting this as unseen-domain evaluation or the exact conditional comparison.

Further prespecified tests also have mixed outcomes. Both TimesFM models fail the benefit-direction criterion, while all six pass a separate layout-selector comparison. Four of 18 risk-accounting endpoints fail (two each for Moirai and Sundial), so that family fails. The fixed response predictor also fails its six-model family: Moirai and Sundial fail against zero response, although all six pass against unit response. Later refits do not revise these failures. The complete supplementary record provides the complete families and follow-ups; the reproduction package's `RESULTS_INDEX.md` identifies their saved outputs and reconstruction scope.