WHEN SHOULD A FORECAST BE CORRECTED? PREDICTION QUALITY IS NOT DECISION QUALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc forecasting systems can leave two predictions for the same future target: a baseline forecast and a proposed revision. The system must decide whether to apply the revision before the target, and hence its true benefit, is known. We study this candidate-conditioned decision as *forecast override*. Under squared error, a revision's realized benefit depends on its alignment with the baseline residual and its magnitude; forecasts with similar errors can therefore favor opposite actions. We evaluate preserve-or-apply policies with Selection Regret, the excess loss relative to choosing the better action in hindsight. On Weather, a residual-based policy has regret 0.01164, compared with 0.03173 for preserving and 0.09048 for applying every revision. On ETTm2, its point-estimate regret is higher than that of the calibration-selected constant action, although the paired interval includes zero. Across forecasting backbones and candidate sources, predictive scores and average correction gains do not consistently rank execution policies. These results motivate evaluating a proposed revision as a decision whose value depends on the candidate and the information available when it is executed.

1 INTRODUCTION

Post-hoc correction starts from a forecast that already exists. Recent systems revise frozen or deployed predictions through learned residuals, output-space adapters, controllers, reconciliation, or other post-processing mechanisms (Liu et al., 2025b; Im & Kwon, 2026; Liang et al., 2026; Kim et al., 2026; Hyndman et al., 2011; Wickramasuriya et al., 2019; Panagiotelis et al., 2021). These methods make correction generation increasingly powerful: they can propose a new forecast after the baseline has been produced. They also create a second scientific object that is easy to miss. Once both the baseline forecast y_0 and a candidate revision c are available, the system must still decide whether the candidate should be executed or whether the original forecast should be preserved.

Existing post-hoc work primarily studies how to construct the revision. PIR learns instance-dependent revisions; COSA adapts frozen outputs through gated residual correction; and δ-Adapter analyzes lightweight post-processing without retraining the deployed backbone (Liu et al., 2025b; Im & Kwon, 2026; Liang et al., 2026). In particular, the squared-error analysis of δ-Adapter already connects average post-processing risk to residual–correction alignment and correction magnitude. The contribution here is different: an already generated correction creates a downstream execution decision requiring separate evaluation. For a particular forecast and candidate, the relevant evidence must support replacing the available forecast rather than preserving it. A revision can be predictable, or even useful on average, without being the right action on every instance.

The ranking of execution policies changes across settings. On ETTm2, a full-residual predictor reaches cosine similarity 0.821 and utility-sign AUROC 0.888, but the preserve/apply policy induced by that prediction has point-estimate Selection Regret 0.00404, compared with 0.00206 for the calibration-selected constant action, and the paired temporal interval crosses zero. Weather reverses the outcome: the same residual policy has regret 0.01164, below both always-preserve (0.03173) and always-apply (0.09048). Learned PIR revisions expose the same separation from another direction: average MSE improves on all three datasets, while 25.2% of ETTm2 revisions and 20.5% of Weather revisions are harmful. A good predictor, a good average revision, and a good instance-level execution rule need not coincide.

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.

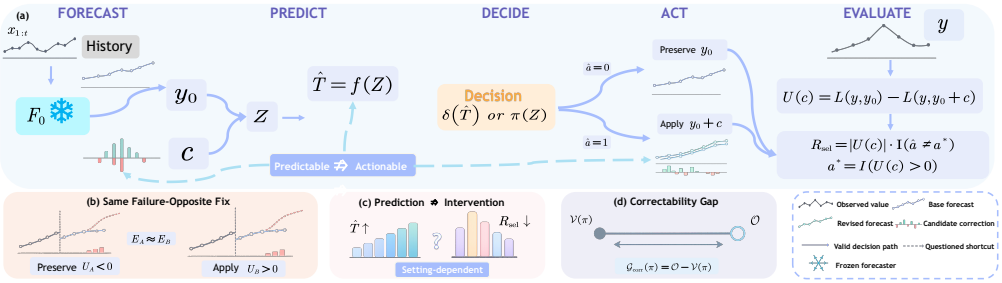

Figure 1: Post-hoc correction as a preserve/apply decision. Prediction supplies inference-time evidence; the decision layer preserves y_0 or applies $y_0 + c$; evaluation uses candidate-conditioned utility and Selection Regret. The lower motifs show same-error/opposite-action cases, prediction–decision mismatch, and unrecovered correction value.

We formalize this conditional execution problem as *forecast override*. Given a baseline forecast y_0, a candidate revision c, and execution-time evidence Z, a policy chooses either to preserve y_0 or output $y_0 + c$. The action is evaluated by the utility of this concrete candidate and by Selection Regret, the loss of the chosen action relative to the better preserve/apply action in hindsight. This lets us ask when inference-time prediction scores recover low-regret execution policies and when they leave value unrecovered through missed benefit or harmful execution.

The experiments follow the discovery path suggested by the formulation. Aggregate failure severity is insufficient because the action also depends on correction direction and scale. Predicting correction-related quantities does not by itself rank execution policies, because the operating rule trades missed benefit against harmful exposure. Changing the backbone, candidate family, or post-hoc system moves the execution boundary rather than removing it. Figure 1 summarizes the setup; the experiments measure the total value left unrecovered by concrete inference-time policies rather than estimating an information-theoretic limit.

2 Related Work

Forecasting backbones. Modern forecasting models span basis-expansion and interpretable sequence architectures (Oreshkin et al., 2020; Lim et al., 2021), efficient and decomposed Transformer variants (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022), and more recent patching, linear, and multiscale designs (Nie et al., 2023; Zeng et al., 2023; Wu et al., 2023; Liu et al., 2024). These works optimize the forecasting map itself. Our scope is conditional on a baseline forecast and a candidate revision already being available.

Post-hoc forecast revision and adaptation. PIR constructs instance-dependent post-hoc revisions (Liu et al., 2025b). COSA directly adapts frozen outputs through gated residual correction, while CTRL decodes controller signals into residual corrections from a frozen backbone (Im & Kwon, 2026; Kim et al., 2026). Liang et al. propose δ-Adapter for post-processing deployed forecasters without retraining and derive squared-error conditions in which residual–correction alignment must offset correction magnitude (Liang et al., 2026). TimeFuse addresses sample-level heterogeneity by learning instance-dependent fusion weights over forecasting models (Liu et al., 2025c). Forecast reconciliation modifies base forecasts through optimal combination, projection, or end-to-end coherence constraints (Hyndman et al., 2011; Wickramasuriya et al., 2019; Panagiotelis et al., 2021; Rangapuram et al., 2021). These works establish that useful post-hoc revisions can be generated. We condition on such a candidate and foreground a different object: the consequence of executing or rejecting that candidate on an individual forecast. Accordingly, our direction–scale identity is used as a candidate-level characterization, not as a claim of originating the alignment–magnitude tradeoff.

Prediction for decisions. Selective prediction and learning to defer study rejection, coverage, and routing decisions (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019; Madras et al., 2018; Mozannar & Sontag, 2020). Unlike selective prediction and deferral, forecast override does not

decide whether a prediction should be trusted or rejected. It decides whether an already available forecast should be replaced by a concrete alternative generated by a correction system. Calibration provides a related notion of predictive reliability (Guo et al., 2017), while recent work on selective-classification evaluation shows that conclusions can depend on the operating criterion used to score abstention systems (Traub et al., 2024). Decision-focused learning and prescriptive analytics show more generally that predictive accuracy need not align with downstream value (Donti et al., 2017; Wilder et al., 2019; Vlastelica et al., 2020; Elmachtoub & Grigas, 2022; Bertsimas & Kallus, 2020). Related work studies goal-oriented forecasting, excess decision risk, actionable calibration, and the value of prediction (Fechete et al., 2026; Perez-Lebel et al., 2025; Rossellini et al., 2025; Fischer-Abaigar et al., 2025). The forecast-specific structure here is a null action that keeps an already available forecast and a candidate-conditioned utility that evaluates one concrete displacement against that null.

3 Forecast Override as a Decision

3.1 Preserve or Apply

Forecast override is a conditional decision object. A post-hoc system has already produced a baseline forecast y_0 and a candidate c. Let Z denote the evidence available when execution is decided. In our experiments, Z may include y_0, c, history summaries, and prior-forecast summaries available at execution time, but it does not include the future target y. A target-based pipeline estimates $\hat{T} = f(Z)$ and applies $\hat{a} = \delta(\hat{T}) \in \{0, 1\}$, where 0 preserves the forecast and 1 applies the candidate. A direct policy uses $\hat{a} = \pi(Z)$; target-based correction is the special case $\pi = \delta \circ f$. The candidate may come from a fixed rule or an instance-dependent learned corrector, abstractly $c = c(Z_c)$. Once proposed, it defines the alternative $y_0 + c$. Appendix A extends the formulation to finitely many candidates.

3.2 Direction–Scale Geometry

Let the forecast be $y_0 \in \mathbb{R}^d$, candidate correction be $c \in \mathbb{R}^d$, and future target be y. Define baseline-relative utility

$$U(c) = L(y, y_0) - L(y, y_0 + c). \tag{1}$$

Positive utility favors applying the candidate. Under MSE, with $e = y - y_0$,

$$U(c) = \frac{1}{d}\left(\|e\|^2 - \|e - c\|^2\right) \tag{2}$$

$$= \frac{1}{d}\left(2\langle e, c\rangle - \|c\|^2\right). \tag{3}$$

Let $E = \|e\|^2/d$, $C = \|c\|^2/d$, $\rho = \langle e, c\rangle/(\|e\|\|c\|)$, and $r = \sqrt{C/E}$. For $E, C > 0$,

$$\Gamma(c) = \frac{U(c)}{E} = 2\rho r - r^2, \qquad \Gamma(c) > 0 \iff 2\rho > r. \tag{4}$$

Here E measures baseline failure severity. The sign of candidate utility also depends on (ρ, r): whether the proposed move aligns with the residual and whether its magnitude is appropriate relative to that residual. Equation (4) is the per-instance action form of the same alignment–magnitude tradeoff that appears in analyses of output-space post-processing (Liang et al., 2026). We use it to characterize a candidate-specific execution boundary rather than to claim a new general squared-error identity.

3.3 Selection Regret and Recoverable Value

For $a \in \{0, 1\}$, write $L_a = L(y, y_0 + ac)$. The better action in hindsight is $a^* = \mathbf{1}[U(c) > 0]$. A policy $\hat{a} = \pi(Z)$ incurs

$$R_{\text{sel}} = L_{\hat{a}} - \min(L_0, L_1) = |U(c)|\,\mathbf{1}[\hat{a} \neq a^*]. \tag{5}$$

For direct utility prediction with $\hat{a} = \mathbf{1}[\hat{U} > 0]$, every wrong action satisfies $R_{\text{sel}} \leq |U - \hat{U}|$. Utility error thus controls decision regret pointwise; correlation, cosine similarity, and AUROC do not provide the same guarantee.

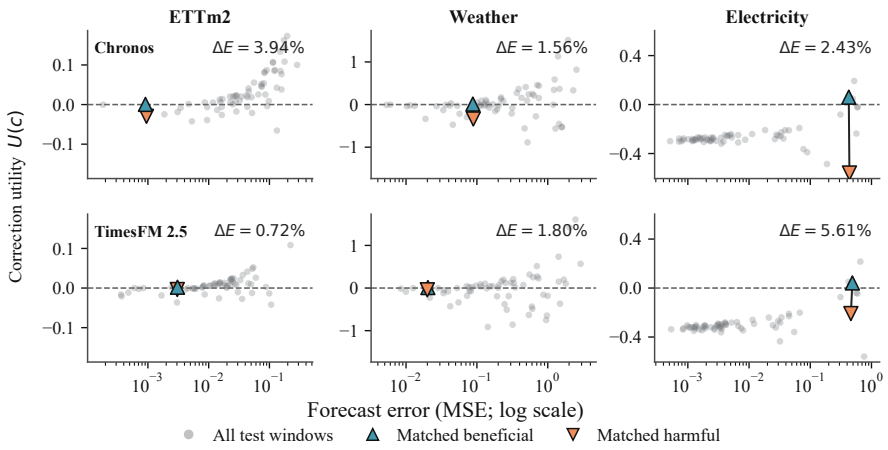

Figure 2: Nearly equal forecast errors can require opposite actions. Gray points are test windows; triangles mark the nearest pair with opposite utility signs. Utility above zero favors applying the candidate, while utility below zero favors preserving the baseline.

Let $U_+ = \max(U, 0)$. Selection Regret decomposes as

$$R_{\text{sel}} = \underbrace{(1 - \pi(Z))U_+}_{\text{missed opportunity}} + \underbrace{\pi(Z)(-U)_+}_{\text{harmful exposure}}. \tag{6}$$

With $\mathcal{O} = \mathbb{E}[U_+]$ and $\mathcal{V}(\pi) = \mathbb{E}[\pi(Z)U]$,

$$\mathcal{G}_{\text{corr}}(\pi) = \mathcal{O} - \mathcal{V}(\pi) = \mathbb{E}[R_{\text{sel}}]. \tag{7}$$

Let $\mathcal{A}_Z = \mathbb{E}[(\mathbb{E}[U \mid Z])_+]$ denote the value of the Bayes preserve/apply policy available from evidence Z. Then

$$\mathcal{G}_{\text{corr}}(\pi) = \underbrace{\mathcal{O} - \mathcal{A}_Z}_{\text{Bayes evidence gap}} + \underbrace{\mathcal{A}_Z - \mathcal{V}(\pi)}_{\text{policy gap}}. \tag{8}$$

This is a population identity. Our experiments observe $\mathcal{O}$ and the value of fitted policies, but do not estimate the unrestricted Bayes value $\mathcal{A}_Z$; they therefore cannot empirically assign the total gap to these two terms or establish fundamental non-recoverability. Appendix A gives the identity, nested-information result, and proofs.

4 Experimental Protocol

We first use Chronos-Bolt tiny (Ansari et al., 2024) on ETTm2 (OT, h=24), Weather (T(degC), h=48), and Electricity (channel 0, h=48). Four candidate-related targets are learned with the same inference-safe evidence and MLP capacity: the full residual e, direct utility U, alignment $A = d^{-1}\langle e, c\rangle$, and directional projection $s = d^{-1}\langle e, q\rangle$ with $q = c/(\|c\|^2 + \epsilon)^{1/2}$. Each prediction is converted to the same preserve/apply utility boundary and evaluated on 64 non-overlapping test windows per setting. Candidate construction and learner family are held fixed, so the matched comparison changes only the prediction target.

We then change one part of the pipeline at a time. TimesFM 2.5 (Das et al., 2024) covers 14 dataset–horizon settings and three correction families. Moirai 2.0 Small (Liu et al., 2025a) replaces the forecasting backbone on three representative settings while retaining the generic-residual downstream pipeline. PIR with PatchTST (Nie et al., 2023) supplies learned, instance-dependent revisions. We additionally evaluate official COSA PatchTST outputs on pre-specified ETTm2 and Weather h=96 settings without modifying COSA's adapter, gating, or optimization logic. The matched target learners are also repeated over five pre-specified initializations. For the primary Chronos, TimesFM, and Moirai analyses, calibration and test windows follow verified chronological separation, correction-utility labels are generated with purged blocked out-of-fold fitting, and test futures are never used for model, threshold, or policy selection. Paired uncertainty uses 2,000 resamples for the primary moving-block validation, with circular block lengths up to 16. PIR exports omit timestamps, so

Table 1: Selection regret comparison for matched target-induced policies; lower is better. Both fixed constant actions are shown. The calibration-selected constant repeats the action chosen using calibration only; the other constant remains a pre-defined held-out reference rather than a test-selected policy.

Policy	ETTm2	Weather	Electricity
Preserve all	0.03066	0.03173	0.04316
Apply all	**0.00206**	0.09048	**0.03546**
Cal.-selected constant	0.00206	0.09048	0.04316
Full residual	0.00404	**0.01164**	0.04079
Direct utility	0.00439	0.01662	0.04074
Alignment	0.00479	0.02987	0.04406
Projection	0.00447	0.03328	0.05174

Table 2: Correction decision boundary examples for the pre-existing near-equal-error pairs. The examples were fixed before the boundary analysis; positive margin $2\rho - r$ favors applying the candidate.

Dataset	Action	E	ρ	r	$2\rho - r$	U
ETTm2	Preserve	0.019154	-0.867	0.487	-2.220	-0.020725
ETTm2	Apply	0.019328	0.982	0.999	0.966	0.018644
Weather	Apply	0.157568	0.956	1.308	0.604	0.124532
Weather	Preserve	0.158375	-0.975	0.217	-2.166	-0.074311

we verified that their rows follow the chronological official test-loader order; the source and export checks are in Appendix B. Selection Regret is the primary decision metric. Full sample counts, learner details, uncertainty procedures, and leakage controls are in Appendix B.

5 Understanding When Corrections Should Be Executed

5.1 Why Forecast Error Alone Cannot Determine Execution

Forecast-error magnitude is informative about utility, but it does not determine the action. True base-forecast MSE has utility-sign AUROC 0.979, 0.959, and 0.769 on ETTm2, Weather, and Electricity, respectively. Under the generic-residual candidate, the pre-existing ETTm2 and Weather examples have nearly equal forecast errors but opposite utilities. Table 2 shows why: one member can cross the execution boundary by pointing against the realized residual, while the other remains well aligned and appropriately scaled.

Across all primary windows, harmful revisions arise through both mechanisms in Equation (4): wrong direction accounts for 46.2%, 61.2%, and 46.3% of harmful cases on ETTm2, Weather, and Electricity, with the remainder positively aligned but overscaled (Appendix Table 19). Thus both direction and relative scale are empirically active rather than merely algebraic terms.

This is not specific to Chronos. All 14 TimesFM settings contain an opposite-utility pair, with 13 of the 14 nearest pairs below the pre-specified 5% relative error-mismatch criterion. Moirai satisfies the same criterion in all three matched settings; the nearest mismatches are 0.441%, 0.044%, and 3.939% on ETTm2, Weather, and Electricity. Figure 2 therefore establishes the narrow empirical point needed here: scalar failure severity can be nearly unchanged while the preferred action flips.

5.2 Can Prediction Recover the Execution Boundary?

Table 1 gives the matched target comparison using absolute Selection Regret. On ETTm2, full-residual prediction reaches cosine 0.821 and utility-sign AUROC 0.888, but its Selection Regret is 0.00404, compared with 0.00206 for the calibration-selected constant action. Direct utility reaches AUROC 0.878 with regret 0.00439; alignment reaches $R^2 = 0.641$ and Pearson 0.816 with still

larger regret. These policies outperform always-preserve on ETTm2 but not the calibration-selected always-apply action, so strong predictive metrics do not rank the available execution rules correctly.

Weather provides the clearest positive regime. Residual and direct-utility policies have regrets 0.01164 and 0.01662, respectively, below both always-preserve (0.03173) and always-apply (0.09048). For the residual policy, the paired difference from preserve-all is -0.02009 with block-16 CI $[-0.03235, -0.00759]$; the difference from apply-all is also negative. The decomposition explains the contrast. Always applying on Weather creates 0.09048 harmful exposure, whereas the residual policy applies 20.3% of candidates, reduces harmful exposure to 0.00389, and misses 0.00775 of available benefit. On ETTm2, selective execution cuts harmful exposure to 0.00103 but gives up 0.00301 of beneficial corrections, so its point-estimate regret exceeds the apply-all constant even though the paired temporal interval crosses zero. Electricity gives the complementary baseline lesson: calibration selects preservation at 0.04316, but the pre-defined apply-all reference has lower held-out regret, 0.03546, than any learned target policy. We report that unselected constant as a reference analysis rather than a test-selected deployment choice.

Additional policy baselines clarify, but do not settle, the threshold explanation. A $|U|$-weighted classifier reduces ETTm2 regret to 0.00232, close to the selected apply-all constant at 0.00206; its paired difference from apply-all crosses zero under both block-1 and block-16 resampling. The same auxiliary grid gives regrets 0.02975 on Weather and 0.05060 on Electricity. Its three-seed direct-utility regressor is not the frozen Table 1 target learner, which is why the Weather and Electricity values differ. A calibration-suffix threshold check gives the same lesson with a different split: tuned thresholds slightly improve ETTm2 residual and direct-utility policies, but do not close the apply-all gap; Weather remains the clearest positive regime, and direct-utility scores there improve sharply after threshold selection (Appendix Tables 6 and 7).

Stability across fitting and temporal resampling. The ETTm2/Weather contrast is not driven by one favorable learner initialization. Across five pre-specified seeds, the residual, direct-utility, and alignment policies improve on the calibration-selected constant in $0/5$ ETTm2 runs; projection improves in $1/5$. On Weather, all four policies improve in $5/5$ runs. The mean regret differences remain positive on ETTm2 and negative on Weather (Appendix Table 8).

The distinction also survives contiguous-window resampling. With circular block length 16, the Weather intervals for residual and utility policies remain below zero, $[-0.09980, -0.05525]$ and $[-0.09756, -0.04774]$, respectively. The corresponding ETTm2 and Electricity intervals cross zero (Appendix Table 9). Thus the strongest positive regime survives both initialization and short-range temporal resampling, while the weaker regimes remain uncertain or unfavorable.

5.3 Does the Phenomenon Generalize Beyond Candidate Design?

TimesFM and Moirai preserve the heterogeneity rather than producing a uniform adverse pattern. On TimesFM ETTm2/h=48, residual cosine is about 0.693 and utility AUROC about 0.783, yet neither induced policy improves on the selected constant; Weather/h=96 is a positive case. On Moirai Weather/h=48, direct utility reduces Selection Regret from 0.15878 to 0.08540 (paired difference -0.07338, 95% CI $[-0.15162, -0.00892]$). The matched ETTm2/h=24 setting is essentially tied with the constant action (0.00223 versus 0.00226).

Electricity/h=48 shows why ranking quality and execution should not be conflated. A residual-derived Moirai score reaches utility-sign AUROC 0.928 and reduces regret from the calibration-selected always-apply action's 0.30606 to 0.02265. Preserving the baseline is still better at 0.01295. The score contains useful decision information, but the policy induced from it does not dominate the available null action.

Figure 3 places the cross-backbone results on the same prediction–decision plane. The horizontal axis is direct-utility sign AUROC and the vertical axis is $R_{\text{global}} - R_{\text{utility}}$. Similar AUROC values occur with both positive and negative decision advantage.

Candidate-family sensitivity. Changing the candidate does not produce a single ordering of decision policies. The cross-correction grid changes candidate construction and, for TimesFM, the source-window grid used to build forecasts. Within that grid, every ETTm2 95% interval for $R_{\text{global}} - R_{\text{utility}}$ includes zero across both Chronos and TimesFM and across residual, polyno-

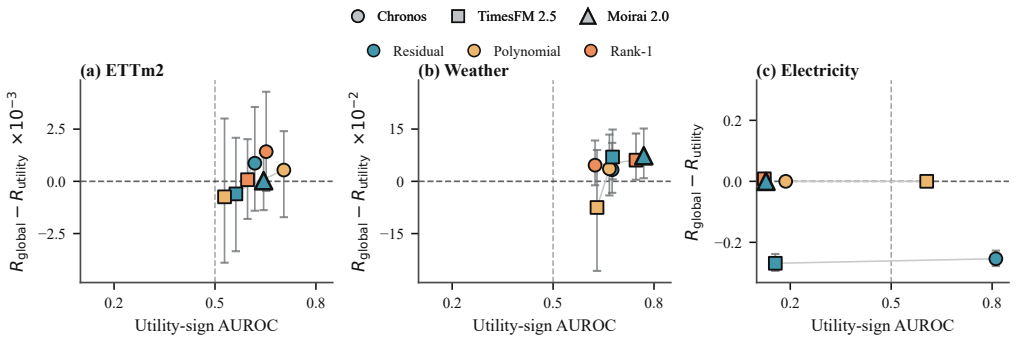

Figure 3: Predictive quality and override value across model and correction families. The horizontal axis is direct-utility sign AUROC; the vertical axis is $R_{\text{global}} - R_{\text{utility}}$, so positive values favor the learned utility policy. Vertical bars are paired 95% intervals for decision advantage. Marker shape denotes backbone and color denotes correction family. Moirai appears only for the generic-residual candidate evaluated on the three representative settings.

Table 3: Independent COSA validation on a real post-hoc correction system. ΔMSE is COSA minus base MSE; negative values favor COSA. The interval is a 95% paired circular moving-block interval with block length 16.

Dataset	ΔMSE	Harmful (%)	Native R_{sel}	Preserve R_{sel}	Block-16 95% CI
ETTm2	$+0.000036$	49.7	0.001814	0.001778	$[-0.000371,\ 0.000456]$
Weather	-0.003719	39.1	0.001240	0.004959	$[-0.004675,\ -0.002738]$

mial, and rank-1 candidates. Weather is more selective: TimesFM residual and rank-1 candidates have positive decision advantage, 0.06985 with CI $[0.00565, 0.1488]$ and 0.06082 with CI $[0.00462, 0.1373]$, while the TimesFM polynomial point estimate is negative and uncertain. On Electricity, residual candidates produce strongly negative utility-policy advantage for both Chronos and TimesFM, whereas the polynomial candidate yields no difference. Candidate construction therefore changes the decision boundary rather than acting as a nuisance that can be averaged away (Appendix Table 13).

Independent validation on a real post-hoc correction system. Official COSA outputs provide an independent check that the execution object is not created by our generic-residual construction. Table 3 shows two regimes. On Weather, COSA improves average MSE under block-16 resampling while 39.1% of realized revisions are harmful. On ETTm2, the aggregate difference is essentially zero and beneficial/harmful revisions are nearly balanced. A real frozen-output adapter therefore exposes candidate revisions whose aggregate quality and instance-level utility are distinct.

Learned candidates. PIR provides a separate test in which the candidate itself is learned. Across $11{,}425$, $10{,}444$, and $5{,}165$ windows, mean MSE improves from 0.2018 to 0.1758 on ETTm2, from 0.1941 to 0.1679 on Weather, and from 0.7491 to 0.1801 on Electricity. Yet 25.2% of ETTm2 revisions and 20.5% of Weather revisions increase error. On Electricity, 99.8% are beneficial. A harmful tail does not imply that a gate can identify it: on chronologically ordered held-out tails, the best learned gates improve the native always-apply point estimate by only 0.00051 on ETTm2 and 0.00172 on Weather, and both moving-block intervals cross zero. Every held-out Electricity revision is beneficial, so always-apply is already the appropriate action there. Full PIR and gating results are in Appendix I.

6 Discussion and Limitations

Forecast override is most relevant when deployment or method design already supplies a baseline forecast and a candidate revision as separately available objects, as in frozen-output adaptation, post-hoc revision, reconciliation, or externally supplied corrections. If a system can be retrained end to end from the same information, the preserve/apply interface can be absorbed into a new forecasting

map. In the post-hoc setting studied here, preservation remains a real null action whose consequence can be evaluated without changing how the candidate was generated.

The experiments expose a gap between ranking a candidate and choosing an action. AUROC can be high even when the operating rule has higher regret than preservation, as in Moirai Electricity. Residual cosine measures a different object again: how well residual direction is reconstructed. Average corrected MSE answers another question because gains on many windows can coexist with a harmful tail. Selection Regret evaluates the actual preserve/apply consequence and weights a wrong choice by its realized cost.

Candidate construction is part of that consequence. The cross-correction results show that the same backbone can move between favorable, neutral, and unfavorable execution regimes as the proposed revision changes, while the same candidate family can behave differently across datasets and backbones. The COSA validation sharpens the same point externally: Weather shows a robust aggregate improvement despite a 39.1% harmful-revision tail, whereas ETTm2 is near neutral. This is why candidate-specific utility, rather than forecast error or aggregate gain alone, is the relevant action object. The direction–scale identity provides an exact MSE characterization of that object, and the correction decision boundary analysis shows both misdirection and overscaling in practice.

Equation (8) separates the evidence and policy gaps. Our experiments measure their sum for fitted policies, but do not identify the two components separately. The nested-evidence reference analysis in the appendix therefore probes finite-policy recoverability, not an information-theoretic ceiling.

The evidence has bounded scope. The direction–scale identity is specific to squared error; the primary target study uses 64 non-overlapping test blocks per setting; the five-seed analysis does not vary the test period; PIR artifacts omit timestamps even though their sliding-window test order is verified by source provenance; and the COSA validation covers only two pre-specified PatchTST/h=96 settings without training a new authorization policy. Within that scope, two further TSFM families, several candidate constructions, moving-block comparisons, large learned-revision streams, and an independent closest-work validation support the same empirical claim: prediction quality or aggregate correction quality alone does not consistently determine the lowest-regret execution rule for a concrete candidate revision.

7 Conclusion

Post-hoc forecasting systems can leave a decision after a revision has been produced: execute it, or keep the forecast already available. Conditional on a baseline and candidate already existing, the action depends on candidate-specific utility, not only on how wrong the baseline forecast is. Across three TSFM families, several correction constructions, and independent PIR/COSA post-hoc outputs, predictive evidence and aggregate gains support low-regret execution in some regimes but not consistently across the tested settings. Treating the revision as an action makes that boundary explicit.

Reproducibility statement

The experimental protocol, target definitions, leakage controls, uncertainty procedures, robustness analyses, and full supplementary results are documented in the appendix. Model, threshold, and policy selection use calibration data only; test futures are reserved for final evaluation and explicitly retrospective quantities.

AI use statement

Generative AI tools were used only to assist with literature lookup, citation cross-checking, and language polishing. They were not used to design experiments, change numerical results, alter methods, or make core scientific claims. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

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.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020. doi: 10.1287/mnsc.2018.3253.

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.

Priya Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Luca-Andrei Fechete, Mohamed Sana, Fadhel Ayed, Nicola Piovesan, Wenjie Li, Antonio De Domenico, and Tareq Si Salem. Goal-oriented time-series forecasting: Foundation framework design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21065–21073, 2026. doi: 10.1609/aaai.v40i25.39249.

Unai Fischer-Abaigar, Christoph Kern, and Juan Carlos Perdomo. The value of prediction in identifying the worst-off. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17239–17261, 2025.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Rob J. Hyndman, Roman A. Ahmed, George Athanasopoulos, and Han Lin Shang. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9): 2579–2589, 2011. doi: 10.1016/j.csda.2011.03.006.

Jeonghwan Im and Hyuk-Yoon Kwon. COSA: Context-aware output-space adapter for test-time adaptation in time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=L7Z5wBMPrW.

Minkyoung Kim, Daeun Ji, Yohan Lee, Beomsoo Kim, and Beakcheol Jang. CTRL: Control-based time series forecasting with LLM-guided residual learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21952–21968. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1104.

Daojun Liang, Qi Li, Yinglong Wang, Jing Chen, Hu Zhang, Xiaoxiao Cui, Qizheng Wang, and Shuo Li. The forecast after the forecast: A post-processing shift in time series. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6f7d90b1198fec96defd80b5ebd5bc81-Abstract-Conference.html.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Doyen Sahoo, Caiming Xiong, Silvio Savarese, 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*, 2024.

Zhiding Liu, Mingyue Cheng, Guanhao Zhao, Jiqian Yang, Qi Liu, and Enhong Chen. Improving time series forecasting via instance-aware post-hoc revision. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Zhining Liu, Ze Yang, Xiao Lin, Ruizhong Qiu, Tianxin Wei, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Breaking silos: Adaptive model fusion unlocks better time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40022–40042, 2025c.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020.

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

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

Anastasios Panagiotelis, Puwasala Gamakumara, George Athanasopoulos, and Rob J. Hyndman. Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting*, 37(1):343–359, 2021. doi: 10.1016/j.ijforecast.2020.06.004.

Alexandre Perez-Lebel, Gael Varoquaux, Sanmi Koyejo, Matthieu Doutreligne, and Marine Le Morvan. Decision from suboptimal classifiers: Excess risk pre- and post-calibration. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2395–2403, 2025.

Syama Sundar Rangapuram, Lucien D. Werner, Konstantinos Benidis, Pedro Mercado, Jan Gasthaus, and Tim Januschowski. End-to-end learning of coherent probabilistic forecasts for hierarchical time series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8832–8843, 2021.

Raphael Rossellini, Jake A. Soloff, Rina Foygel Barber, Zhimei Ren, and Rebecca Willett. Can a calibration metric be both testable and actionable? In *Proceedings of the Thirty-Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4937–4972, 2025.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 2323–2347, 2024.

Marin Vlastelica, Anselm Paulus, Vít Musil, Georg Martius, and Michal Rolinek. Differentiation of blackbox combinatorial solvers. In *International Conference on Learning Representations*, 2020.

Shanika L. Wickramasuriya, George Athanasopoulos, and Rob J. Hyndman. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019. doi: 10.1080/01621459.2018.1448825.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, 2021.

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.

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.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286, 2022.

A Proofs and Formal Extensions

A.1 Utility-prediction control of Selection Regret

Proposition S1. For the direct-utility policy $\hat{a} = \mathbf{1}[\hat{U} > 0]$ and oracle action $a^* = \mathbf{1}[U > 0]$,

$$R_{\text{sel}} = |U|\mathbf{1}[\hat{a} \neq a^*] \leq |U - \hat{U}|. \tag{9}$$

Proof. When $\hat{a} = a^*$, the left side is zero. Otherwise U and $\hat{U}$ lie on opposite sides of zero, so the interval joining them contains zero and $|U| \leq |U - \hat{U}|$. Taking expectations gives $\mathbb{E}[R_{\text{sel}}] \leq \mathbb{E}|U - \hat{U}|$. Utility MAE therefore provides pointwise decision control, whereas correlation, cosine similarity, and AUROC do not provide the same guarantee.

A.2 Population evidence gap and nested information

Proposition S2. Let

$$\begin{aligned} \mathcal{O} &= \mathbb{E}[U_+], & \mathcal{A}_Z &= \mathbb{E}\big[(\mathbb{E}[U \mid Z])_+\big], \\ P_Z &= \mathbb{E}[U_+ \mid Z], & N_Z &= \mathbb{E}[(-U)_+ \mid Z]. \end{aligned} \tag{10}$$

Then

$$\mathcal{O} - \mathcal{A}_Z = \mathbb{E}[\min(P_Z, N_Z)]. \tag{11}$$

If $\sigma(Z_1) \subseteq \sigma(Z_2)$, then $\mathcal{A}_{Z_1} \leq \mathcal{A}_{Z_2}$.

Proof. Since $\mathbb{E}[U \mid Z] = P_Z - N_Z$,

$$\mathcal{O} - \mathcal{A}_Z = \mathbb{E}[P_Z - (P_Z - N_Z)_+] = \mathbb{E}[\min(P_Z, N_Z)]. \tag{12}$$

For nested information, the tower property and convexity of $(\cdot)_+$ yield

$$\mathcal{A}_{Z_1} = \mathbb{E}\big[(\mathbb{E}[\mathbb{E}[U \mid Z_2] \mid Z_1])_+\big] \leq \mathcal{A}_{Z_2}. \tag{13}$$

This is a population statement. The experiments in this paper do not estimate the unrestricted Bayes value $\mathcal{A}_Z$. Held-out values from a finite, validation-selected policy family need not be monotone because estimation and model-selection error remain. The nested-evidence experiment therefore probes finite-policy recoverability; it is not an estimator of $\mathcal{O} - \mathcal{A}_Z$ and cannot establish an information-theoretic ceiling.

A.3 Finite candidate action spaces

Let $c_0 = 0$ denote preservation and $c_1, \ldots, c_K$ denote proposed revisions. Define $U_k = L(y, y_0) - L(y, y_0 + c_k)$, with $U_0 = 0$. The evidence-conditioned Bayes action selects $\arg\max_k \mathbb{E}[U_k \mid Z]$. Candidate generation, candidate ranking, and execution authorization therefore remain distinct in the multi-candidate case.

A.4 Execution cost and randomized authorization

Suppose executing a candidate incurs a nonnegative cost κ in the same utility units as the loss reduction. Define net utility

$$U_\kappa(c) = L(y, y_0) - L(y, y_0 + c) - \kappa = U(c) - \kappa. \tag{14}$$

The hindsight action applies the candidate iff $U(c) > \kappa$, and all binary Selection-Regret formulas carry over after replacing U by U_κ. Thus the zero-utility boundary used in the main experiments is the special case $\kappa = 0$; asymmetric deployment costs shift the execution threshold without changing the decision object.

Throughout the main paper, $\pi(Z) \in \{0, 1\}$ denotes a deterministic hard action. If a system instead outputs an application probability $p(Z) \in [0, 1]$ and samples the action, the conditional expected regret at $\kappa = 0$ is

$$\mathbb{E}[R_{\text{sel}} \mid U, Z] = (1 - p(Z))U_+ + p(Z)(-U)_+. \tag{15}$$

Equation (6) is the deterministic special case $p(Z) = \pi(Z)$.

B EXPERIMENTAL PROTOCOL

Table 4 collects the experimental grids and their comparability boundaries. They are not pooled into a single leaderboard: each changes a different part of the pipeline. The primary grid holds the backbone, candidate construction, evidence, and learner capacity fixed while changing the prediction target. TimesFM and Moirai change the forecasting backbone; the candidate-family grid changes the proposed revision; PIR replaces the hand-constructed candidate with a learned instance-dependent one. Rows with similar names across tables should therefore be compared within their own grid unless the table explicitly states that all other factors are held fixed.

Table 4: Experimental grid map.

Grid ID	Settings and size	What changes	Held fixed / comparability boundary
G1 Primary Chronos	ETTm2 h=24 ($N_{\text{cal}} = 512$, $N_{\text{test}} = 64$); Weather h=48 (256, 64); Electricity h=48 (128, 64)	Prediction target: residual, utility, alignment, projection	Chronos-Bolt tiny, generic-residual candidate, `history+prior+c` evidence, MLP $(64, 32)$, and 64 non-overlapping test windows; isolates target-to-decision behavior.
G2 Policy baselines	Same three G1 settings and evidence	Policy family: utility regression, unweighted utility-sign classification, $\lvert U\rvert$-weighted utility-sign classification	Same chronological split and input variant as G1; a policy-baseline check, not an additional target row in Table 1.
G2b Threshold check	Same three G1 settings and evidence; train/threshold suffix sizes are $410/102$, $205/51$, and $102/26$ for ETTm2, Weather, and Electricity	Scalar execution threshold selected on a later calibration suffix	Calibration/test separation uses the G1 raw-interval overlap purge; the prefix/suffix split adds no extra fixed-length purge. Threshold candidates are preserve-all, apply-all, zero, observed calibration scores, and adjacent-score midpoints; ties minimize calibration regret, then correction rate, then threshold value.
G3 TimesFM 2.5	14 dataset–horizon settings over ETTh1, ETTh2, ETTm1, ETTm2, Weather, Exchange, Electricity	Forecast backbone and dataset–horizon setting	Generic-residual downstream pipeline and decision definitions; broad cross-backbone comparison.
G4 Moirai 2.0	Three representative settings using the same 35% source-window grid as G3	Forecast backbone only	Generic-residual candidate, targets, seed mapping, and decision analysis remain fixed relative to the TimesFM confirmation.
G5 Candidate families	ETTm2, Weather, Electricity representative settings under Chronos and TimesFM	Candidate construction: residual, polynomial, rank-1	Evaluates candidate-conditioned decision boundaries. Because the candidate and, for TimesFM, the source-window grid change, G5 values are compared within G5 rather than treated as direct G1 repeats.
G6 PIR	ETTm2, Weather, Electricity full exported PatchTST/PIR streams	Candidate source becomes a learned instance-dependent revision	Same consequence metrics; tests whether average learned-revision gains eliminate harmful instances or make authorization trivial.

Grid ID	Settings and size	What changes	Held fixed / comparability boundary
G7 COSA	Official PatchTST outputs on ETTm2/Weather h=96	Candidate comes from an independent frozen-output test-time adapter	No COSA adapter, gate, optimizer, or scientific logic is modified; tests whether the execution object exists in closest work.

B.1 TARGETS AND DECISION RULES

All matched target learners use the same inference-safe evidence and MLP body. Full-residual prediction estimates $\hat{e}$ and reconstructs

$$\hat{U} = d^{-1}\big(2\langle \hat{e}, c\rangle - \|c\|^2\big)\,. \tag{16}$$

Direct utility estimates $\hat{U}$ directly. Alignment predicts $A = d^{-1}\langle e, c\rangle$ and uses $\hat{U} = 2\hat{A} - C$, where $C = d^{-1}\|c\|^2$. Directional projection predicts $s = d^{-1}\langle e, q\rangle$ for $q = c/\sqrt{\|c\|^2 + \epsilon}$ and reconstructs the same alignment. Target-induced policies use $\hat{a} = \mathbf{1}[\hat{U} > 0]$ with no test-tuned threshold. The policy-baseline grid in Table 6 adds two classifiers on the same inference-safe evidence: an unweighted utility-sign classifier and a regret-aligned classifier that predicts $\mathbf{1}[U > 0]$ with sample weight $|U|$.

Table 6: Additional same-evidence policy baselines on the primary Chronos grid. All learned policies use the inference-safe `history+prior+c` evidence, hidden layers $(64, 32)$, ReLU activations, Adam, three seeds, and calibration-only fitting. Lower Selection Regret is better. Each learned-policy row averages the three seeded model scores before applying the fixed decision threshold and computing Selection Regret. This is an auxiliary policy-baseline run; Table 1 reports the frozen primary target learner, so its direct-utility row need not match this table.

Policy baseline	ETTm2	Weather	Electricity
Preserve all	0.03066	0.03173	0.04316
Apply all	0.00206	0.09048	0.03546
Direct utility regression	0.00439	0.02197	0.05362
Unweighted utility-sign classifier	0.00653	0.02544	0.04493
$\lvert U\rvert$-weighted utility-sign classifier	0.00232	0.02975	0.05060

The policy-baseline grid uses the same chronological splits and input variant as the primary target learners, but it is a policy comparison rather than another row of Table 1. It checks how alternative training objectives behave under the same evidence. Weighting by $|U|$ helps on ETTm2 relative to the unweighted classifier and direct utility regression, but the result is still close to the selected apply-all constant. Its paired difference from apply-all is $+0.00026$, with block-1 interval $[-0.00065, 0.00131]$ and block-16 interval $[-0.00084, 0.00123]$. On Weather, the residual and direct-utility target policies in the primary grid remain the strongest reported policies; the regret-aligned classifier does not reproduce their low regret. On Electricity, all three learned policy baselines remain worse than the better fixed action.

Table 7 isolates the zero-threshold question. For this check, each score model is trained on the earlier calibration prefix; a scalar threshold is selected on the later chronological calibration suffix; and the test set is used only after the threshold is frozen. Because the score learner sees less calibration data than the primary Table 1 learners, these rows test threshold sensitivity rather than replacing the primary target rows. On ETTm2, calibration-selected thresholds slightly improve residual and direct-utility policies, but both remain worse than the apply-all constant by point estimate. On Weather, the residual zero-threshold policy remains below preserve-all under block lengths 4, 8, and 16, and below apply-all under all blocks; direct-utility scores benefit strongly from threshold selection. Electricity remains a boundary case, where direct-utility thresholding improves the point estimate but the paired interval crosses zero.

Table 7: Calibration-selected threshold checks on the primary Chronos grid. Lower Selection Regret is better; negative Δ favors the first policy.

Dataset	Score	$R_{\tau=0}$	τ_{cal}	$R_{\tau_{\text{cal}}}$	$\Delta(\tau_{\text{cal}} - 0)$, block-16 CI
ETTm2	Residual	0.00338	-0.00210	0.00336	-0.00003 $[-0.00008, 0.00000]$
ETTm2	Direct utility	0.00412	-0.00126	0.00380	-0.00032 $[-0.00097, 0.00000]$
Weather	Residual	0.01687	-0.01169	0.01755	$+0.00068$ $[-0.00202, 0.00338]$
Weather	Direct utility	0.07376	$+0.04313$	0.01786	-0.05590 $[-0.08001, -0.03224]$
Electricity	Residual	0.03844	$+0.00012$	0.03844	$+0.00000$ $[0.00000, 0.00000]$
Electricity	Direct utility	0.04102	-0.02110	0.03344	-0.00758 $[-0.02629, 0.00356]$

Key paired fixed-action checks. G1 primary Weather residual vs. preserve-all: $\Delta = -0.02009$, block-1 CI $[-0.03951, -0.00335]$, block-16 CI $[-0.03235, -0.00759]$. G2b Weather residual, $\tau = 0$ vs. preserve-all: $\Delta = -0.01486$, block-1 CI $[-0.03419, 0.00109]$, block-16 CI $[-0.02762, -0.00154]$. G2b Weather residual, $\tau = 0$ vs. apply-all: $\Delta = -0.07361$, block-1 CI $[-0.09821, -0.04756]$, block-16 CI $[-0.09669, -0.04932]$. ETTm2 $|U|$-weighted classifier vs. apply-all: $\Delta = +0.00026$, block-1 CI $[-0.00065, 0.00131]$, block-16 CI $[-0.00084, 0.00123]$.

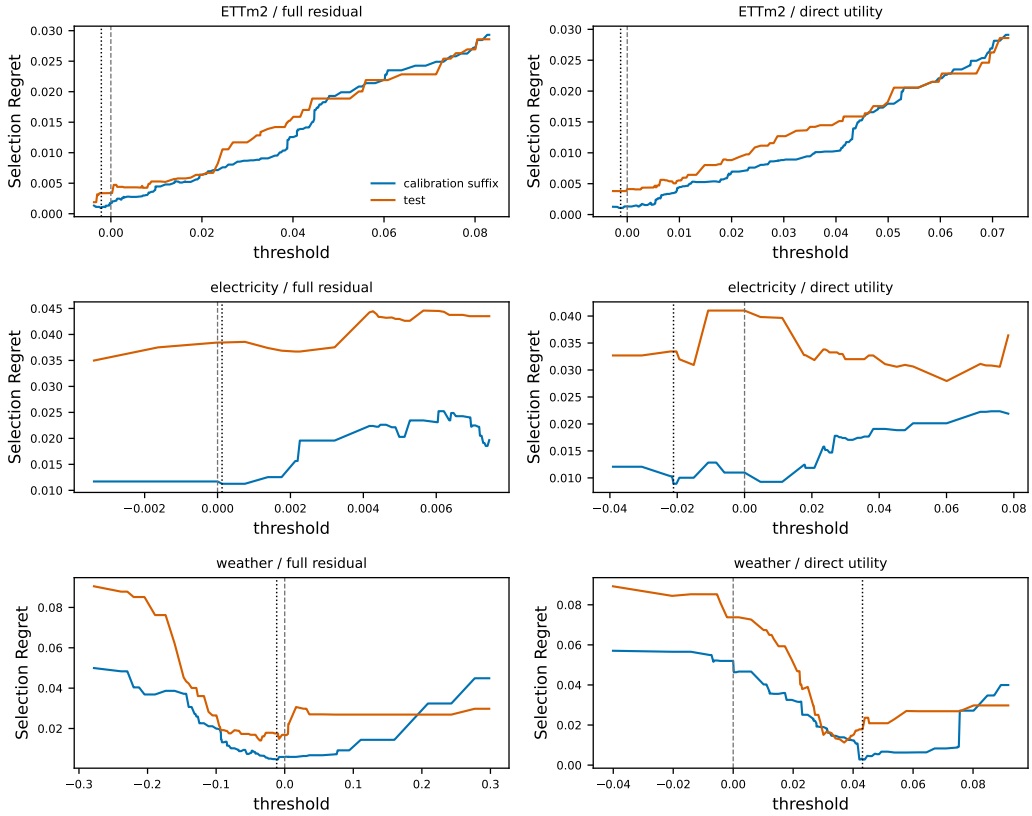

Figure 4: Threshold-regret curves. Dotted lines mark calibration-selected thresholds; dashed lines mark the fixed zero threshold. Test curves are retrospective diagnostics only.

B.2 Chronological separation and OOF labels

For the primary Chronos analysis and the TimesFM/Moirai confirmations, normalization uses calibration-prefix statistics only. Primary stride is 96 and horizons are at most 48, so primary future target intervals do not overlap. Whenever calibration correction targets or utilities train a downstream policy, labels are generated by purged blocked out-of-fold fitting. A labeled window is never

Table 8: Five-seed decision sensitivity of the matched target learners. "Improve" counts seeds with negative regret difference, which favors the target-induced policy.

Target	$R_{\text{target}} - R_{\text{global}}$ mean $\pm$ std	Improve
ETTm2		
Full residual	$0.00192 \pm 4.81\text{e}{-4}$	0/5
Direct utility	$0.00227 \pm 2.03\text{e}{-4}$	0/5
Alignment	$0.00238 \pm 2.21\text{e}{-4}$	0/5
Projection	0.00138 ± 0.00109	1/5
Weather		
Full residual	-0.07348 ± 0.00347	5/5
Direct utility	-0.06856 ± 0.00479	5/5
Alignment	-0.05892 ± 0.00137	5/5
Projection	-0.05959 ± 0.00364	5/5
Electricity		
Full residual	-0.00391 ± 0.00834	4/5
Direct utility	$-1.02\text{e}{-4} \pm 0.00750$	3/5
Alignment	0.00153 ± 0.00454	2/5
Projection	0.00916 ± 0.00182	0/5

evaluated by a corrector trained on that window or an overlapping raw interval. Test futures are used only for final evaluation and explicitly retrospective quantities.

PIR artifacts do not store calendar timestamps, but their test-stream chronology is verified from source provenance. At the frozen official PIR commit, the test DataLoader uses no shuffling; ETTm2 and custom datasets map increasing sample index to increasing sliding-window start within a forward test block; the official test loop appends batches in loader order; and our export-only patch concatenates base forecasts, PIR forecasts, targets, and failure scores in exactly that order. The exact streaming-retrieval patch changes only retrieval execution and the finalizer only copies artifacts, so neither reorders windows. The gate split therefore follows the verified chronological order of the official sliding-window test stream, with purge gaps between train, validation, and held-out tails. The artifact itself still cannot provide human-readable calendar timestamps.

For Electricity, exact retrieval uses chunked computation to reduce memory use while preserving the same model, data, retrieval count, stride, precision, optimization settings, and global top-k rule. Numerical equivalence to the full-bank computation was verified on ETTm2 and Weather with $\text{atol} = \text{rtol} = 10^{-5}$.

C Robustness to Initialization and Temporal Dependence

C.1 Five pre-specified initializations

The matched target learners are rerun with five initialization seeds; the data split, candidate, features, architecture, optimizer, and decision rule are unchanged. This isolates whether the ETTm2/Weather contrast is a consequence of one favorable fit.

The pattern is stable on the two datasets that anchor the main comparison. On Weather, all four target-induced policies improve on the calibration-selected constant in all five runs. On ETTm2, residual, utility, and alignment improve in none of the five runs, while projection improves once. Electricity sits closer to the boundary: residual improves in four runs, utility in three, and alignment in two, with standard deviations comparable to their mean differences. We therefore treat Electricity as an unstable boundary case rather than using its seed average as evidence for a consistent gain.

C.2 Paired temporal-dependence sensitivity

The primary confidence intervals compare paired test-window regret differences. To assess whether short-range temporal dependence makes those intervals artificially narrow, we use $2{,}000$ paired circular moving-block resamples. Block length 1 is the paired-window reference; lengths 4, 8, and

Table 9: Paired Selection-Regret differences under paired-window and block-16 resampling. Negative values favor the target-induced policy.

Data	Policy	Point	Win.	Block 16
ETT	Resid.	0.00198	$[-0.00123, 0.00626]$	$[-0.00138, 0.00536]$
ETT	Util.	0.00233	$[-8.51\text{e}{-4}, 0.00682]$	$[-8.94\text{e}{-4}, 0.00563]$
Wea.	Resid.	-0.07884	$[-0.1024, -0.05537]$	$[-0.09980, -0.05525]$
Wea.	Resid. vs preserve	-0.02009	$[-0.03951, -0.00335]$	$[-0.03235, -0.00759]$
Wea.	Util.	-0.07385	$[-0.09873, -0.04945]$	$[-0.09756, -0.04774]$
Elec.	Resid.	-0.00237	$[-0.02490, 0.01972]$	$[-0.02605, 0.01987]$
Elec.	Util.	-0.00243	$[-0.02278, 0.01872]$	$[-0.02200, 0.01558]$

Table 10: Forecast-override accounting for the matched target-induced policies. Retrospective opportunities are 0.03066, 0.03173, and 0.04316 for ETTm2, Weather, and Electricity; Corr. is the correction rate (%).

Data	Policy	Recov.	Harm	Miss/R_{sel}	Value	Corr.
ETT	Global	0.03066	0.00206	0/0.00206	0.02860	100.0
	Residual	0.02765	0.00103	0.00301/0.00404	0.02661	84.4
	Utility	0.02730	0.00103	0.00336/0.00439	0.02627	81.2
	Alignment	0.02730	0.00144	0.00336/0.00479	0.02586	84.4
	Projection	0.02730	0.00111	0.00336/0.00447	0.02619	82.8
Wea.	Global	0.03173	0.09048	0/0.09048	-0.05875	100.0
	Residual	0.02398	0.00389	0.00775/0.01164	0.02009	20.3
	Utility	0.02173	0.00663	0.01000/0.01662	0.01511	20.3
	Alignment	0.00484	0.00299	0.02689/0.02987	0.00186	7.8
	Projection	0.00541	0.00696	0.02632/0.03328	-0.00155	14.1
Elec.	Global	0	0	0.04316/0.04316	0	0.0
	Residual	0.02242	0.02005	0.02074/0.04079	0.00237	85.9
	Utility	0.01879	0.01636	0.02437/0.04074	0.00243	75.0
	Alignment	0.01657	0.01746	0.02660/0.04406	-0.00089	70.3
	Projection	0.02302	0.03159	0.02015/0.05174	-0.00858	85.9

16 preserve increasingly long local runs. Table 9 shows the reference and the most conservative block-16 intervals.

Weather is insensitive to this change: the residual and utility intervals remain entirely below zero at block length 16. ETTm2 crosses zero under both the paired-window and block-16 analyses, as does Electricity. The resampling analysis therefore strengthens the Weather positive case without converting the weaker settings into significant effects.

D Where Selection Regret Comes From

Selection Regret can arise in two different ways: applying a harmful candidate or preserving the baseline when the candidate would have helped. Table 10 decomposes the matched target policies into these two terms, making clear why similar correction rates can have different decision value.

ETTm2 illustrates the missed-opportunity trade-off. The calibration-selected constant applies every candidate and recovers the full retrospective opportunity, at the cost of 0.00206 harmful exposure. Selective policies roughly halve that harmful exposure but give up 0.00301–0.00336 of beneficial corrections, so their total regret is larger. Weather has the opposite geometry: always applying creates 0.09048 harmful exposure, while always preserving has regret 0.03173; the residual policy applies only 20.3% of candidates and reduces total regret to 0.01164. On Electricity, calibration selects preservation; the residual policy recovers 0.02242 of opportunity but incurs 0.02005 harmful exposure, leaving only a small positive policy value. These decompositions are more informative than correction rate or Regret Closed alone.

Table 11: TimesFM 2.5 results over 14 dataset–horizon settings. Panel A reports predictive reference measures and retrospective opportunity; Panel B reports Selection Regret, correction rate (%), and paired 95% confidence intervals.

Panel A: Predictive reference measures

Setting	$\mathcal{O}$	Cos.	Res. AUC	Util. AUC
ETTh1/h=24	0.01367	0.05566	0.5368	0.4537
ETTh2/h=24	0.03365	0.1709	0.3580	0.6339
ETTm1/h=48	0.02329	0.2809	0.5329	0.5193
ETTm2/h=24	0.01032	0.5124	0.4656	0.5355
ETTm2/h=48	0.1079	0.6934	0.4415	0.7830
Electricity/h=24	0.01340	0.05894	0.2966	0.1379
Electricity/h=48	0.00489	0.06135	0.9333	0.08889
Electricity/h=96	0.00397	0.1149	0.05556	0.08333
Exchange/h=24	0.02773	-0.06789	0.5145	0.5242
Exchange/h=48	0.02518	-0.05843	0.5839	0.4178
Exchange/h=96	0.04654	-0.1036	0.4723	0.5536
Weather/h=24	0.04045	0.1583	0.6286	0.6378
Weather/h=48	0.1256	0.1557	0.6838	0.7172
Weather/h=96	0.1702	0.3058	0.6442	0.6632

Panel B: Decision outcomes

Setting	R_g	R_r	R_u	Corr.	$R_u - R_g$ 95% CI
ETTh1/h=24	0.01367	0.01560	0.01805	42.9	$[-0.002, 0.011]$
ETTh2/h=24	0.03365	0.03503	0.03365	0.0	$[0.00, 0.00]$
ETTm1/h=48	0.02329	0.01707	0.02375	46.0	$[-0.011, 0.013]$
ETTm2/h=24	0.00424	0.00397	0.00478	49.2	$[-0.002, 0.003]$
ETTm2/h=48	0.01592	0.01592	0.01592	100.0	$[0.00, 0.00]$
Electricity/h=24	0.01340	0.2105	0.2065	95.2	$[0.165, 0.218]$
Electricity/h=48	0.00489	0.01032	0.2837	100.0	$[0.250, 0.303]$
Electricity/h=96	0.00397	0.2902	0.2795	93.7	$[0.244, 0.302]$
Exchange/h=24	0.02773	0.02931	0.02773	0.0	$[0.00, 0.00]$
Exchange/h=48	0.02518	0.03057	0.02518	0.0	$[0.00, 0.00]$
Exchange/h=96	0.04654	0.05818	0.03927	27.5	$[-0.040, 0.021]$
Weather/h=24	0.04045	0.02601	0.03279	31.7	$[-0.029, 0.013]$
Weather/h=48	0.1256	0.07712	0.07779	22.2	$[-0.124, 0.011]$
Weather/h=96	0.1702	0.07574	0.08019	46.0	$[-0.193, -0.009]$

E TimesFM 2.5 Across the Extended Grid

The TimesFM grid asks whether the main contrast survives a different forecasting family and a broader set of dataset–horizon combinations. It spans ETTh1, ETTh2, ETTm1, ETTm2, Weather, Exchange, and Electricity. Every setting has positive retrospective opportunity and an opposite-utility pair; $13/14$ nearest pairs satisfy the pre-specified 5% error-mismatch criterion.

The predictive and decision columns do not move together. ETTm2/h=48 has residual cosine 0.6934 and utility AUROC 0.7830, yet the direct-utility policy exactly matches the global-action regret at 0.01592. Electricity/h=48 is sharper: the residual-induced utility-sign AUROC is 0.9333, while the residual-policy regret (0.01032) remains above the global preserve regret (0.00489). High separability is therefore present even though the induced action rule is not the lowest-regret choice.

Weather supplies the complementary case. At h=96, the direct-utility policy reduces regret from 0.1702 to 0.08019, and the paired 95% interval for $R_u - R_g$ is $[-0.193, -0.009]$. The extended grid is thus not a collection of failures: it contains a clear positive regime alongside null and adverse ones.

Table 12: Moirai 2.0 Small on the three matched settings. Residual AUC is the utility-sign AUROC induced by the full-residual predictor; Utility AUC is the direct-utility predictor AUROC. Pair gap is the relative forecast-error mismatch of the nearest opposite-utility pair. Lower Selection Regret is better.

Data	h	$\mathcal{O}$	Resid. cos.	Resid. AUC	Utility AUC	Pair gap (%)
ETTm2	24	0.00949	0.165	0.621	0.644	0.441
Weather	48	0.15878	0.242	0.691	0.770	0.044
Electricity	48	0.01295	0.052	0.928	0.128	3.939

Data	R_{preserve}	R_{global}	R_{residual}	R_{utility}	$R_{\text{utility}} - R_{\text{global}}$ [95% CI]
ETTm2	0.00949	0.00226	0.00313	0.00223	-0.00003 [-0.00137, 0.00138]
Weather	0.15878	0.15878	0.08189	0.08540	-0.07338 [-0.15162, -0.00892]
Electricity	0.01295	0.30606	0.02265	0.30606	0.00000 [0.00000, 0.00000]

F Moirai 2.0 Across Three Representative Settings

Moirai 2.0 Small (Liu et al., 2025a) replaces the forecasting backbone while the generic-residual correction, target learners, decision rules, seed mapping, and Selection-Regret analysis remain fixed. The three settings reuse the same 35% chronological source-window grid as the TimesFM experiment, so the comparison changes the forecasting model without introducing a new downstream search.

The three rows deliberately cover different decision regimes. On ETTm2/h=24, direct utility is indistinguishable from the calibration-selected constant action: Selection Regret is 0.00223 versus 0.00226, and the paired interval crosses zero. On Weather/h=48, direct utility reduces regret from 0.15878 to 0.08540, with paired difference -0.07338 and 95% CI $[-0.15162, -0.00892]$.

Electricity/h=48 separates ranking from execution most clearly. The residual-induced score reaches utility-sign AUROC 0.928 and lowers regret from the calibration-selected always-apply action (0.30606) to 0.02265. Preservation is still better at 0.01295. The direct-utility learner, meanwhile, collapses to the always-apply action and inherits its regret. Thus the residual score contains substantial decision information, but that fact alone does not certify the thresholded action rule.

All three settings also contain an opposite-utility pair within the 5% forecast-error mismatch criterion: 0.441% on ETTm2, 0.044% on Weather, and 3.939% on Electricity. The same-error phenomenon therefore persists after the backbone substitution.

G Candidate-Family Sensitivity

The cross-correction grid tests a direct alternative explanation: perhaps the observed decision behavior is peculiar to the generic-residual candidate. The grid is not a direct replicate of the primary Table 1 comparison. It keeps the three representative dataset–horizon settings but changes candidate construction, and the TimesFM rows use the corresponding TimesFM source-window grid. Table 13 reports $R_g - R_u$, so positive values favor direct-utility authorization within that grid.

Table 13: Direct-utility decision advantage across candidate families. Positive $R_g - R_u$ favors direct-utility authorization; Harm is the harmful-application rate (%).

Data	Correction	TSFM	$R_g - R_u$ [95% CI]	Harm
ETTm2	Residual	Chronos	8.68e-04 [-0.00141, 0.00356]	18.8
		TimesFM	-6.02e-04 [-0.00335, 0.00208]	31.0
	Polynomial	Chronos	5.41e-04 [-0.00172, 0.00240]	20.8
		TimesFM	-7.42e-04 [-0.00390, 0.00301]	35.7
	Rank-1	Chronos	0.00142 [-4.63e-04, 0.00429]	19.2
		TimesFM	7.81e-05 [-0.00180, 0.00202]	30.0
Weather	Residual	Chronos	0.03329 [-0.03305, 0.1105]	40.0

Data	Correction	TSFM	$R_g - R_u$ [95% CI]	Harm
		TimesFM	0.06985 [0.00565, 0.1488]	30.0
	Polynomial	Chronos	0.03600 [-0.04051, 0.1342]	30.0
		TimesFM	-0.07509 [-0.2571, 0.08993]	60.0
	Rank-1	Chronos	0.04633 [-0.01137, 0.1172]	33.3
		TimesFM	0.06082 [0.00462, 0.1373]	21.1
Electricity	Residual	Chronos	-0.2538 [-0.2778, -0.2271]	95.2
		TimesFM	-0.2687 [-0.2937, -0.2382]	96.6
	Polynomial	Chronos	0.0000 [0.0000, 0.0000]	0.0
		TimesFM	0.0000 [0.0000, 0.0000]	0.0
	Rank-1	Chronos	0.0000 [0.0000, 0.0000]	95.2
		TimesFM	0.00820 [0.0000, 0.02461]	93.5

ETTm2 remains close to a null regime: all six confidence intervals include zero. Weather is candidate- and backbone-dependent. Under TimesFM, residual and rank-1 candidates have positive decision advantage with intervals above zero, whereas the polynomial candidate has a negative point estimate with a wide interval. On Electricity, residual candidates behave differently again: direct-utility authorization is much worse than the global action for both Chronos and TimesFM ($R_g - R_u = -0.2538$ and -0.2687). The polynomial candidate yields no difference, while rank-1 is neutral for Chronos and weakly positive for TimesFM.

The candidate family therefore changes the decision boundary rather than removing the prediction–decision distinction. A backbone-level predictive score cannot by itself determine whether an arbitrary proposed revision should be executed.

H Finite-Policy Evidence Richness and Error-Conditioned Action Heterogeneity

Finite validation-selected learners need not inherit the population monotonicity of Proposition S2. Table 14 tests this in two ways: Panel A fits the same policy family with increasingly rich evidence, while Panel B asks how often similar scalar forecast errors still contain opposite optimal actions. Neither panel estimates the unrestricted Bayes value $\mathcal{A}_Z$.

Table 14: Finite-policy evidence reference analysis.

Panel A: Finite-policy recoverability under nested evidence

Data	Evidence	Dim.	Value	R_{sel}	Corr.
ETT	Failure	1	0.02892	0.00173	87.5
	Base	19	0.02803	0.00263	85.9
	Candidate	32	0.02778	0.00287	85.9
	Rich pair	176	0.02791	0.00274	85.9
Wea.	Failure	1	-0.05242	0.08415	96.9
	Base	19	-0.02231	0.05404	40.6
	Candidate	32	-0.01200	0.04373	29.7
	Rich pair	224	-0.03585	0.06758	70.3
Elec.	Failure	1	-0.00214	0.04530	75.0
	Base	19	0.00469	0.03848	67.2
	Candidate	32	0.00153	0.04163	48.4
	Rich pair	224	-0.01747	0.06063	25.0

Panel B: Error-conditioned action heterogeneity

Metric	ETTm2	Weather	Elec.
N	64	64	64
Beneficial (%)	79.7	23.4	35.9
Entropy	0.157	0.307	0.773
Mixed bins (%)	40.0	40.0	100.0
Nearest ΔE (%)	0.897	0.510	0.715
Opp. @ 1%	11.1	14.3	22.2

Metric	ETTm2	Weather	Elec.
Opp. @ 5%	9.1	13.8	27.8
Opp. @ 10%	6.9	20.0	39.4

Panel A is not monotone on held-out data. On Weather, candidate-aware evidence improves on the scalar failure feature, but the highest-dimensional rich-pair representation gives back part of that gain. On Electricity, the base evidence has lower regret than candidate and rich-pair representations. This does not contradict Proposition S2: the proposition concerns the Bayes value available from nested information, whereas the table contains finite-sample fitted policies selected through validation.

Panel B shows why scalar failure severity is an incomplete decision variable. Every dataset contains mixed error bins; Electricity has mixed actions in all bins and the nearest opposite-action pair differs in forecast error by only 0.715%. At a 10% error-neighborhood radius, 39.4% of Electricity neighborhoods contain an opposite action. These statistics establish local action heterogeneity conditional on similar scalar error; they do not by themselves establish irreducible ambiguity given the full inference-time evidence.

I Learned Candidates: PIR and Gating

PIR provides a qualitatively different candidate source: the revision is learned and instance-dependent rather than generated by the generic residual, polynomial, or rank-1 constructions. The first question is whether average improvement removes harmful instances.

Table 16: Full-stream learned-corrector transfer. Native PIR always applies its revision, so missed opportunity is zero and Selection Regret equals harmful exposure.

Metric	ETTm2	Weather	Electricity
Base/PIR MSE	0.2018/0.1758	0.1941/0.1679	0.7491/0.1801
Beneficial/Harmful (%)	74.8/25.2	79.5/20.5	99.8/0.2
Harm exposure	0.01282	0.00595	0.00016

Average MSE improves on all three streams, but the instance-level picture differs sharply. Roughly one quarter of ETTm2 revisions and one fifth of Weather revisions are harmful, whereas Electricity is almost uniformly beneficial. The aggregate gain therefore identifies a useful corrector without implying that every revision should be executed.

Held-out gate validation. For each dataset, evaluation windows are split in verified chronological test-stream order into purged train, validation, and held-out tails; the source-level row-order checks are described in Appendix B. Model family and threshold are fixed on the earlier train/validation portions before held-out evaluation. The gate experiment asks whether the harmful revisions are predictable from the available inference-time evidence. Table 17 reports recovered benefit and harmful exposure (Recov./Harm), correction rate (Corr.), and regret intervals relative to native PIR.

Table 17: Held-out preserve/apply results for PIR.

Policy	Recov./Harm	Missed/R_{sel}	Corr.	95% CI
ETTm2				
Preserve	0/0	0.05553/0.05553	0.0	[0.010,0.059]
Native PIR	0.05553/0.02124	0/0.02124	100.0	[0,0]
Failure gate	0.02291/0.01220	0.03262/0.04482	10.0	[0.015,0.033]
Weighted gate	0.05537/0.02057	1.61e-4/0.02073	98.2	$[-0.001, 0.000]$
Oracle	0.05553/0	0/0	75.0	$[-0.038, -0.010]$
Weather				
Preserve	0/0	0.03481/0.03481	0.0	[0.022,0.036]
Native PIR	0.03481/0.00559	0/0.00559	100.0	[0,0]

Policy	Recov./Harm	Missed/R_{sel}	Corr.	95% CI
Failure gate	0.03255/0.00474	0.00226/0.00700	94.3	$[-0.000, 0.003]$
Weighted gate	0.03466/0.00372	1.50e-4/0.00387	99.2	$[-0.005, 0.000]$
Oracle	0.03481/0	0/0	83.1	$[-0.010, -0.003]$
Electricity				
Preserve	0/0	0.5446/0.5446	0.0	[0.519,0.567]
Native / learned / Oracle	0.5446/0	0/0	100.0	[0,0]

On ETTm2 and Weather, the weighted gate gives the best learned-gate point estimate, but it changes the native correction rate only slightly (98.2% and 99.2%) and both intervals include zero. The failure-only gate is markedly worse on ETTm2. On the Electricity tail, every revision is beneficial, so native always-apply, the learned gate, and the oracle coincide. Harmful revisions are therefore present in two streams, but these finite held-out gate experiments do not provide a reliable improvement over native execution.

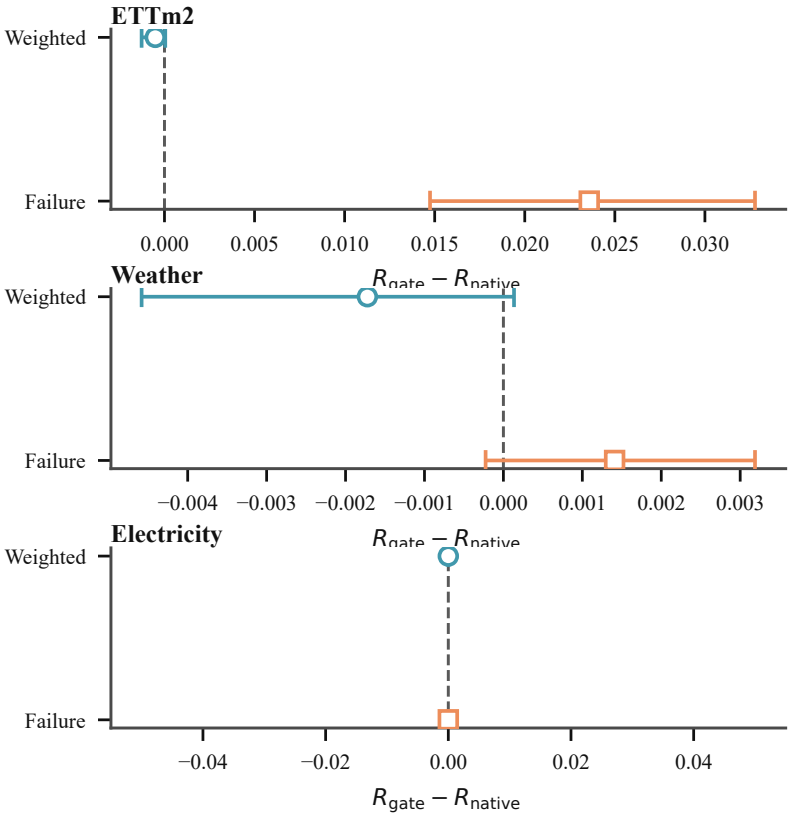

Figure 5: Held-out gate intervals relative to native PIR; values below zero favor the gate.

J Correction Decision Boundary Analysis

Main Table 2 gives the worked direction–scale mechanism analysis for the fixed boundary examples. Here we report the full primary-window boundary analysis rather than repeat that table. Recomputing (ρ, r) for every primary test window shows two distinct routes to harmful execution: a candidate can point against the realized residual ($\rho \leq 0$), or it can retain positive alignment while exceeding the beneficial scale boundary ($r \geq 2\rho$). Both mechanisms occur in every primary setting, with dataset-dependent prevalence.

Table 19: Mechanism analysis for harmful generic-residual revisions crossing the MSE execution boundary. "Wrong dir." means $\rho \leq 0$; "overscale" means $\rho > 0$ but $r \geq 2\rho$. The last two columns compare median geometry for beneficial/harmful revisions.

Dataset	Harmful	Wrong dir. (%)	Overscale (%)	ρ med. (+/−)	r med. (+/−)
ETTm2	13/64	46.2	53.8	0.992 / 0.156	0.924 / 2.615
Weather	49/64	61.2	38.8	0.956 / −0.246	0.957 / 2.873
Electricity	41/64	46.3	53.7	0.565 / 0.038	0.332 / 0.756

K Independent Validation on a Real Post-hoc Correction System

To test whether the preserve/apply object is peculiar to our hand-constructed candidates, we evaluate predictions exported by the official COSA implementation (Im & Kwon, 2026) at frozen commit `527c0fe`, using PatchTST with horizon 96 on ETTm2 and Weather. No COSA gate, adapter, or optimization rule is modified. We define the realized candidate as $c = y_{\mathrm{COSA}} - y_0$ and evaluate the same baseline-relative utility and consequence quantities used elsewhere in the paper. Main Table 3 exposes the core positive/near-neutral boundary; the fuller validation below retains sample counts and raw base/COSA MSE values.

Table 20: Independent validation using official COSA PatchTST predictions at horizon 96. ΔMSE is COSA minus base MSE; negative values favor COSA. The final column is a 95% paired circular moving-block interval (block length 16) for ΔMSE. Selection Regret treats the native COSA output as always executing the realized revision $c = y_{\mathrm{COSA}} - y_0$.

Dataset	N	Base/COSA MSE	ΔMSE	Ben./Harm. (%)	Native R_{sel}	Preserve R_{sel}	Block-16 95% CI
ETTm2	13,841	0.159586/0.159622	+0.000036	50.3/49.7	0.001814	0.001778	[−0.000371, 0.000456]
Weather	10,444	0.172041/0.168322	−0.003719	60.9/39.1	0.001240	0.004959	[−0.004675, −0.002738]

The two settings form a useful boundary pair. On Weather, COSA improves average MSE by 0.003719 and the block-16 interval remains entirely below zero, while 39.1% of individual realized revisions are harmful. The aggregate gain is therefore genuine, but it does not imply uniformly positive per-window utility. On ETTm2, average COSA performance is essentially tied with the base forecast: the MSE difference is $+3.6 \times 10^{-5}$ and every reported paired interval crosses zero, while beneficial and harmful revisions occur at nearly equal rates. These results are not evidence that COSA fails; they show that a real frozen-output adaptation system naturally exposes the same candidate-execution evaluation object in both positive and near-neutral regimes.