THE PRICE OF CHOOSING: FORECAST SELECTION UNDER FRESH CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When a forecast chosen with one calibration pool is calibrated again before deployment, the two calibrations need not rank candidates alike. We make the expected interval score after a fresh, finite calibration pool the comparison target. An exact construction shows that reusing one pool can select the worse forecast regardless of the number of queries. For A fixed predictors and N observations, rank weights and two hinges compute the complete-subset risk estimate and every fixed-pool-size delete-one value exactly in $O(AN \log N)$ arithmetic operations; across 12 settings, this is 1.44–24.69 times faster than a custom batched dense reference computing the same outputs. For globally nearby endpoints with a common bounded scale, calibration ranks give a paired score envelope without bounding the response, and its uniform constant 8 is sharp across ranks. Label-cost bounds then separate exact assessment from reliable selection. In a 72-setting unbounded-support study, the envelope enables beneficial controlled updates in up to 20 settings, while a query-level bound never updates. On nine public forecasting datasets, calibration removes most of a matched fitting gain, from 5.34 to 1.52 percent, and from 7.05 to 3.13 percent over ten seeds; in a prespecified external test, the seven hourly datasets, one of its two primary groups, retain 62 percent of the gain. Pre-calibration comparisons can therefore overstate the surviving difference. Candidate forecasts should be compared by their risk after the calibration they will actually receive, and its exact estimate is now cheap to compute.

1 INTRODUCTION

Suppose two weather predictors will each have their prediction intervals adjusted using the most recent observations before tomorrow's interval is issued. This adjustment is called calibration. Which predictor should be deployed? Evaluating candidates with one historical calibration pool measures performance conditional on that pool, whereas deployment uses a new one, and collecting more evaluation outcomes against the same pool does not remove this discrepancy. The relevant object is the expected score of the forecast *after a fresh, finite calibration*.

For example, with four calibration observations, a reused pool without the rare outcome favors an aggressive interval over a moderate one (scores about 1.75 versus 1.83). Repeating calibration reverses their order: expected scores become about 2.63 versus 1.83 (Table 1 gives exact values). More test queries against the same reused pool cannot reveal this reversal.

The issue arises whenever candidates are compared on one calibrated validation pool and the chosen forecast is calibrated again on newer data before deployment. Our target charges both width and missed outcomes, unlike coverage-based selection. We hold fitted candidates and pool size fixed and average over the calibration that follows selection. Section 6 relates this target to coverage guarantees, forecast-comparison tests and cross-validation.

Contributions. First, an exact construction shows that reusing one calibration pool can select the worse forecast however many queries are made against it, and the ordering extends to bounded continuous perturbations (Section 2). Second, we compute the complete-subset risk estimate and *all* fixed-m delete-one estimates exactly in $O(AN \log N)$ arithmetic operations. Deleting an observation removes both its query and calibration roles while leaving the predictor and deployment pool size unchanged, so the output supports exact assessment and sensitivity analysis without enumerating

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.

subsets or recomputing the statistic N times. In twelve matched settings, a custom batched dense reference computing the same outputs takes 1.44–24.69 times as long (Sections 3 and 5). Third, for globally nearby endpoints with a common bounded scale, a paired score envelope removes the query-level $1/\alpha$ penalty even when responses are unbounded. Its uniform constant 8 is sharp across ranks, and standard concentration then controls harmful selection (Section 4). Fourth, label-cost bounds separate exact assessment from reliable selection. Along $\alpha = 1/(m+1)$, the necessary labels grow as $(m+1)(\epsilon/\gamma)^2$ for a small benefit γ at endpoint distance ϵ; at fixed 90 percent coverage, a specified atomic family requires $\Theta(m \log(1/\delta))$ labels for sufficiently large m and a fixed risk gap, an order that independent blocks attain. Fifth, experiments show that the envelope enables beneficial controlled updates in up to 20 of 72 unbounded-support settings, where a query-level bound never updates, and that calibration removes most of a matched fitting gain on nine public forecasting datasets, from 5.34 to 1.52 percent, and from 7.05 to 3.13 percent over ten seeds (Section 5 and Appendix M.1).

The practical message is direct: compare candidate forecasts by their risk after the calibration they will actually receive. The complete-subset estimate of that risk, together with every delete-one value, is now exact and cheap to compute.

2 The Risk of a Fresh Calibration

Condition on A fitted predictors with ordered endpoints $L_a(x), U_a(x)$ and positive scales $c_a(x)$. Fitting and hyperparameter choice precede selection. A size-m pool calibrates candidate a using residuals $r_{a,i} = \max\{L_a(X_i) - Y_i, Y_i - U_a(X_i)\}/c_a(X_i)$ and $q_a = r_{a,(k)}$, where $k = \lceil (m+1)(1-\alpha) \rceil$. The issued interval is $[L_a(x) - c_a(x)q_a, U_a(x) + c_a(x)q_a]$. Signed correction allows negative q_a; nonnegative correction replaces it by $\max(q_a, 0)$. Inverted endpoints collapse to their original midpoint. If $k > m$, the correction is infinite, not the largest observed residual.

For ordered finite endpoints define the interval score (Gneiting & Raftery, 2007)

$$S_\alpha([L, U], y) = U - L + \frac{2}{\alpha} \max\{L - y, y - U, 0\}.$$

For IID observations $Z = (X, Y)$ from a target distribution P, let

$$R_m(a) = \mathbb{E}_{C_m \sim P^m, Z \sim P} S_\alpha(I_{a,C_m}(X), Y), \qquad \mathcal{R}(\widehat{a}) = \mathbb{E}_D R_m(\widehat{a}(D)) - \min_a R_m(a). \quad (1)$$

Selection data D, fitting, final pool C_m and final test data have separate roles, and risks are finite. A realized pool's conditional risk differs from R_m, which averages over the pool itself.

	Keep	Moderate	Aggressive
One reused pool	2	$11/6$	$\mathbf{7/4}$
Fresh calibration pool	2	$\mathbf{11/6}$	$18191/6912$

Table 1: The lowest expected interval score changes when calibration is repeated. In this exact example (Section 2.1), $m = 4$, $\alpha = .2$, and the reused pool has no rare observation. Bold marks the preferred forecast in each row.

2.1 Why more queries may not help

Take a rare state of probability $p = 1/[2(m+2)]$, outcome 1 in that state and -1 otherwise. The unchanged forecast is $[-1, 1]$; a moderate forecast is $[0, 2]$ or $[-2, 2]$ in the rare or common state; an aggressive forecast is $[0, 0]$ or $[-2, 1]$. With unit scales and $\alpha = 1/(m+1)$, a pool missing the rare state has conditional risks $(2, 2-2p, 2-3p)$. The aggressive forecast wins strictly. Its fresh risk instead exceeds the moderate forecast's by $\Delta_m = 3 - p - 3(1-p)^m > 0$. The pool misses the rare state with probability $(1-p)^m$, regardless of the number of queries.

Proposition 6 gives a strictly positive finite-query regret bound and extends the ordering to bounded continuous perturbations, whose law and admissible noise radius depend on m. The error of fixed-pool assessment therefore does not vanish as queries accumulate. Averaging over calibration pools targets the fresh risk directly, and the next section makes that average exact and fast.

3 Exact Assessment Without Enumerating Subsets

Let $d = m + 1$. On a d-point block, let K_a average the d scores obtained by making each point the query and using the others for calibration. The complete U-statistic over $N \geq d$ selection observations has mean $R_m(a)$. Equivalently, average every query's score over size-m subsets of the other $N - 1$ observations. Overlapping queries are not independent replications.

Sort the residuals once. After excluding query j, the probability that position r supplies the correction is

$$w_r = \frac{\binom{r-1}{k-1}\binom{N-1-r}{m-k}}{\binom{N-1}{m}}.$$

Tied residuals retain their observation indices; the weights count subsets of indices. For a query with center z, half-width v, scale c and $e = |y - z|$, the score at correction t is exactly

$$\ell(t) = \frac{2e}{\alpha} + (2 - 2/\alpha)c(t + v/c)_+ + (2/\alpha)c\{t - (e - v)/c\}_+. \tag{2}$$

This includes midpoint collapse. Two weighted prefix sums evaluate each hinge above its breakpoint; excluding a query shifts the weights to its right.

Proposition 1 (Complete risk and all deletions)**.** *For fixed $m, 1 \leq k \leq m, 0 < \alpha < 1$ and A fitted candidates on the same $N \geq m + 2$ finite observations, the complete risk estimates and all N delete-one estimates per candidate can be computed in $O(AN \log N)$ arithmetic operations and $O(N)$ workspace per candidate. Each deletion removes both observation roles, preserving m, k and the fitted predictor. Signed and nonnegative correction, ties and midpoint collapse are included. Candidates may have different positive scales, each shared by its own two tails. Retaining all output vectors uses $O(AN)$ storage. The computation does not require IID observations.*

For deletion i, the remaining query averages over $N-2$ possible calibration observations. Two shifted rank-weight arrays and the query's subtraction reuse the same hinge prefix sums (Appendix A.1). The same pass computes paired jackknife quantities for $O(A)$ predetermined comparisons, and Appendix A.2 discusses classical inference from the deletion values. Algorithm 1 uses this computation for its complete-U contrasts. Because the output is the exact complete-subset statistic, whose mean is the fresh risk $R_m(a)$, the right comparison is available at quasi-linear cost; the next section adds the uncertainty bound that a controlled decision needs.

4 A Sharp Paired Envelope and the Label Cost of Selection

Compare two fixed predictors a, b sharing the same positive scale $c(x) \leq c_{\max}$, with a known bound on the entire input support:

$$\max\{|L_a(x) - L_b(x)|, |U_a(x) - U_b(x)|\} \leq \epsilon c(x). \tag{3}$$

For example, apply a bounded correction $\epsilon c(x)h(x)$, $\|h(x)\|_\infty \leq 1$, to baseline endpoints and project onto ordered pairs; a standard bounded-output head enforces this geometry by design. If also $c \geq c_{\min} > 0$ and the response and baseline endpoints have finite first moments, fresh risks are finite, while the response may still have unbounded support.

Proposition 2 (A sharp uniform paired envelope)**.** *Under equation 3, with $d = m + 1$, $k \leq m$ and $s = d - k$, every block satisfies*

$$|K_a - K_b| \leq B_{ab} := c_{\max}\epsilon \min\{4/\alpha,\ 4 + 4s/(\alpha d)\} \leq 8c_{\max}\epsilon. \tag{4}$$

The bound includes ties, signed or nonnegative correction and midpoint collapse. No uniform constant smaller than 8 works over either class across all finite ranks.

Why the penalty does not explode. Residual and quantile differences are at most ϵ, so issued endpoints differ by at most $2c(x)\epsilon$. Each candidate misses at most s of the d leave-one-out queries. Positive paired distance terms can occur only in one candidate's missed set, negative terms only in the other's. The signed sum therefore has magnitude at most $2sc_{\max}\epsilon$. Adding width and penalty differences gives equation 4. A construction with strictly positive corrections attains $8\epsilon(d-1)/d$ in both classes; see Appendix B.

Algorithm 1 Compare fitted forecasts under fresh calibration

Require: Fitted candidates and reference b; common scale and certified global endpoint bounds; N independent selection labels; final size m, level α and error budget δ.

1: **if** there is no distinct eligible candidate **then**
2: Keep b.
3: **else**
4: Compute eligible complete-U contrasts with the rank-prefix formula using $d = m + 1$.
5: Use the tightest fit-fixed geometric envelope available to every method; form confidence bounds with a joint error budget across candidates and tests.
6: Among candidates with negative upper bounds choose the smallest, breaking ties by a fixed order; if none passes, keep b.
7: **end if**
8: Only now draw the independent size-m final calibration pool and issue the chosen forecast's interval.

Removing signed radius changes. Signed calibration cancels a candidate-specific constant shift of its standardized raw radius. Write z_a, v_a for center and half-width, and set $u(x) = |z_a - z_b|/c$ and $w(x) = (v_a - v_b)/c$. If the global suprema $A_+ = \sup_x(u + w)$ and $A_- = \sup_x(u - w)$ are finite, the same comparison has the valid effective radius

$$\epsilon_{ab}^{\text{eff}} = \frac{A_+ + A_-}{2} = \inf_t \sup_x \{u(x) + |w(x) + t|\}. \tag{5}$$

The bound in Proposition 2 holds with this radius under signed correction, as a consequence of radius-shift invariance (Appendix B).

A controlled comparison. For N IID selection observations from the same law defining R_m, independent of fitting, let $\widehat{D}_{ab}$ be their complete-U contrast. The standard U-Hoeffding argument (Hoeffding, 1963) gives, with $n = \lfloor N/d \rfloor$,

$$\Pr\{R_m(a) - R_m(b) > \widehat{D}_{ab} + t\} \le \exp\{-nt^2/(2B_{ab}^2)\}.$$

For $A - 1$ comparisons, use $t_{ab} = B_{ab}\sqrt{2\log((A-1)/\delta)/n}$ and update only when $\widehat{D}_{ab} + t_{ab} < 0$. The probability of choosing a candidate with worse fresh mean risk is at most δ. This is standard risk testing (Angelopoulos et al., 2025), and variance-aware U bounds can use the same envelope (Peel et al., 2010). Algorithm 1 states the inputs, the decision and the independent final-calibration step.

The cost of detecting a benefit. For one candidate with true fresh-risk advantage $\gamma > 0$, the same concentration bound shows that an update is selected with probability at least $1 - \beta$ if

$$\gamma > B_{ab}\left\{\sqrt{\frac{2\log(1/\delta)}{n}} + \sqrt{\frac{2\log(1/\beta)}{n}}\right\}.$$

For $\delta = \beta = .05$, using the universal $B_{ab} = 8c_{\max}\epsilon$ is sufficient when $n > 512\log(20)(c_{\max}\epsilon/\gamma)^2$. The next two results give necessary label costs in an explicit family.

Proposition 3 (A necessary number of informative labels)**.** *Let $a = [\epsilon, \epsilon]$, $b = [0, 0]$, $c = 1$, $M > 2\epsilon > 0$, $d = m + 1$ and $\alpha = 1/d$. For $Y \in \{0, M\}$ with rare probability p,*

$$R_m(a) - R_m(b) = -2\epsilon + 4\epsilon(1 - dp)(1 - p)^{d-1}.$$

Any two-action rule whose harmful-update probability is at most δ under $p = 0$ fails to choose the beneficial action under $p = 1/d$ with probability at least $(1 - \delta)(1 - 1/d)^N$. Thus failure at most $\beta < 1 - \delta$ requires

$$N \ge \left\lceil \frac{\log[\beta/(1-\delta)]}{\log(1 - 1/d)} \right\rceil.$$

All labels informative about p, including auxiliary fit or pilot labels, are counted in N; the final calibration pool is not seen during selection.

Table 2: Time to compute complete risks, all fixed-m deletions and the paired jackknife. Ranges span four setting medians, each from three single-threaded calls. Ratio is batched dense time divided by prefix time.

N	Prefix (ms)	Batched dense (ms)	Ratio
512	8.39–9.65	12.24–13.99	1.44–1.46
2,048	34.43–34.64	209.60–213.57	6.07–6.19
8,192	144.52–144.92	3566.63–3576.21	24.65–24.69

The proof applies a classical two-point argument to the indistinguishable all-zero sample, and it holds even when both methods know the support. Along $\alpha = 1/d$, the necessary cost is order $d\log((1-\delta)/\beta)$: the sharp envelope removes the extra query-level penalty, but calibration still sets the scale of the information cost.

Corollary 1 (Gap-dependent cost for two fixed actions)**.** *In the family of Proposition 3, for every $0 < \gamma \le \epsilon/8$ there are $p_- < p_+$ such that the risk difference is γ and $-\gamma$, respectively. Let $0 < \delta, \beta < 1$ and $\delta < 1 - \beta$. If a rule using N independent informative labels selects the harmful action with probability at most δ under p_-, and misses the beneficial action with probability at most β under p_+, then*

$$N \ge \frac{d\epsilon^2}{32\gamma^2} \max\{\mathrm{kl}(\delta, 1-\beta), \mathrm{kl}(1-\beta, \delta)\}.$$

Every label about the population is charged to N; the final pool is not available to the rule.

With a classical change-of-measure step, the explicit fresh-calibration family shows that, at fixed error levels, the necessary labels grow as $d(\epsilon/\gamma)^2$ when the benefit is small. The sufficient bound above has the same dependence in this restricted comparison of two fixed actions along $\alpha = 1/d$.

The cost at fixed coverage. At fixed $\alpha = .1$, two fixed common-scale predictors can require $\Theta(m\log(1/\delta))$ labels to distinguish their fresh risks, for sufficiently large m and $0 < \delta \le 1/4$ (Proposition 10), under either calibration operator at zero raw halfwidth. Free covariates have identical laws in both worlds. Independent-block averaging attains the order. These atomic laws have residual-quantile jumps. In contrast, let both residual quantiles be globally L_a, L_0-Lipschitz, the score be L-Lipschitz in its correction, and every paired query loss have absolute bound B_q. For fixed independent-fit candidates and $N \ge m + 1$ IID selection observations, Efron–Stein gives

$$\mathrm{Var}(U_N) \le \frac{\{2B_q + L(L_a + L_0)\sqrt{2m/(m+1)}\}^2}{2N}.$$

These known class bounds control the full statistic, including higher-order terms (Appendix N.6). A local extension allows unbounded residuals, with an explicit bad-event cost away from extreme ranks.

5 Experiments: Computation, Power and Forecast Quality

We evaluate computational cost, controlled comparison power and fitted forecasting quality in turn.

5.1 Complete risk and all deletions against a batched reference

Both implementations return two complete risks, both full delete-one vectors and their paired jackknife quantity. Our custom dense reference reuses one query-by-threshold loss matrix across all deletions: $O(AN^2)$ arithmetic and storage, rather than N full recomputations. Prefix evaluation uses $O(AN\log N)$ work and $O(N)$ workspace per candidate. The twelve settings cross three N values, $m = 9, 19$, and signed/nonnegative correction at $\alpha = .1$.

Before timing, complete vectors agree to 3.26×10^{-13}, and 306 rational cases agree with literal subset enumeration. Table 2 summarizes all 72 timed calls. Each call uses one CPU task and one numerical-library thread (the allocation permits two logical CPU IDs); calls alternate method order, include weights, sorting, allocation and pairing, and exclude fitting, imports and worker startup. The

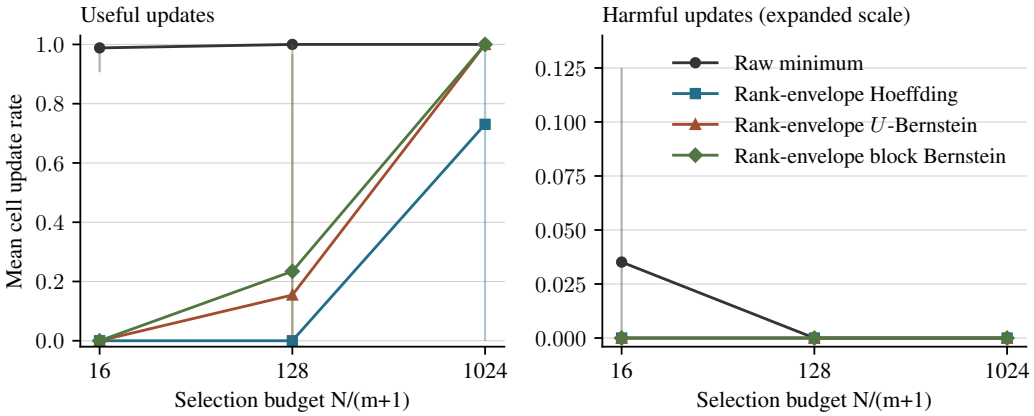

Figure 1: Update frequency by normalized selection budget $N/(m+1)$ over all 72 settings. An independent evaluation reference classifies 16 settings per budget as beneficial and eight as harmful, each with 32 paired repetitions; the harmful-update axis is expanded and vertical marks span settings.

ratio grows with N, from 1.44–1.46 at $N = 512$ to 24.65–24.69 at $N = 8192$, where measured process memory is 56.5–58.2 MiB for prefix versus 657.9–658.0 MiB for dense, including interpreter overhead. Appendix A.1 gives all twelve rows, independent checks and memory measurement details; Appendix A adds separate mean-only comparisons.

5.2 A COMPLETE UNBOUNDED-SUPPORT COMPARISON

We fix two analytic interval predictors differing by a bounded nonlinear center correction. Gaussian and untruncated Student-t_3 noise provide finite risks without a finite response range. Three signal levels, $m = 9, 19$ with $\alpha = 1/(m+1)$, two calibration operators and $N/(m+1) \in \{16, 128, 1024\}$ give 72 settings, each with 32 independent selection repetitions. Candidates, seeds, budgets, tie rules and comparisons are fixed before execution, and the same geometry is available to every rule.

Seven rules use the same candidates and final calibration. Raw and complete-U minimization have no harm guarantee. Three Hoeffding or Bernstein rules compare complete-U means; a further empirical Bernstein control compares independent block means. The U-Bernstein implementation uses a conservative paired-block variance bound rather than the complete order-$2d$ variance estimator of Peel et al. (2010). All adaptive rules consult N selection labels; keep consults none. Candidate fitting costs zero in this analytic comparison, and final-pool labels and true law parameters are used only for evaluation. Appendix B gives formulas, role separation and all costs.

Table 3: Selection outcomes across 72 unbounded-support settings and $2304 = 72 \times 32$ repetitions. Better and worse count settings whose adjusted 95 percent mean-score interval, over 432 contrasts, lies below or above zero.

Rule	Better /72	Worse /72	Updates /2304
Keep	0	0	0
Raw minimum	48	0	1539
Complete-U minimum	48	0	1535
Query-envelope Hoeffding	0	0	0
Rank-envelope Hoeffding	12	0	374
Rank-envelope U-Bernstein	18	0	591
Rank-envelope block Bernstein	20	0	632

The query-envelope rule never updates. The rank envelope makes controlled comparison nonvacuous (Table 3): Hoeffding has 12 favorable settings, complete-U Bernstein 18, and block Bernstein 20, none has an unfavorable adjusted interval, and the confidence rules make zero observed harmful

updates (Figure 1). Raw and complete-U minimization reach 48 favorable settings each but carry no harm guarantee: on settings whose independent evaluation reference resolves the alternative as harmful, they update in up to $4/32$ and $5/32$ repetitions, respectively. With the same labels and geometry, the rank envelope thus supports controlled updates where the query-level bound always abstains.

5.3 Calibration removes most of a fitting gain

The forecasting studies ask what part of a fitted improvement survives final calibration, on trained predictors and dependent time series. Local and amortized fits share 52,986 parameters, initialization, decoder and training settings. Each of three folds withholds three of the nine public datasets. A four-dimensional coefficient is fitted to history inside a Student-t decoder. Each held-out series supplies 40 earlier calibration and 12 later evaluation origins, with signed joint calibration at nominal 80 percent coverage.

Over 153 series and three seeds, local fitting improves the raw interval score by 5.34 percent, but after calibration the gain is 1.52 percent $[.85, 2.18]$. Over ten seeds, every final gain is positive: the raw gain is 7.05 percent and the final gain 3.13 percent (seed-level interval $[1.80, 4.46]$; Appendix M.1). The primary series-weighted final-score difference is $-.1029$ $[-.1522, -.0557]$. In this comparison the surviving gain is narrower intervals: the width component is $-.1178$, while the outside-distance component is $+.0149$ $[-.0159, +.0476]$. Before calibration, the contrast is carried instead by fewer misses at essentially equal width. Coverage is $.8120$ versus $.8146$, with difference interval $[-.0059, .0007]$. A comparison made before calibration would have overstated the gain that reaches the issued intervals. Appendix M gives per-seed results and further forecasting comparisons.

A separately specified external test repeats the matched comparison on held-out datasets of a fixed benchmark snapshot. Write R and F for the candidate-minus-control interval-score contrasts before and after calibration, after context-median and robust-scale normalization, and retention for $|F|/|R|$. On seven hourly datasets, the local advantage shrinks from $R = -.11444$ to $F = -.07101$, a retention of $.6205$ $[.4843, .8162]$ with series resampled within each fixed dataset. Appendix M.6 gives every dataset, both uncertainty calculations and weighting sensitivities.

6 Related Work

The comparison target. Conformalized quantile regression calibrates fitted endpoints (Romano et al., 2019); model selection already has coverage and set-size guarantees (Yang & Kuchibhotla, 2025; Liang et al., 2026; Hegazy et al., 2025). Local selection with safe indexing and ensembling retains marginal coverage (Wang & Wang, 2026), while learned conditional-coverage assessment ranks procedures by deviation from the target level (Zhou et al., 2026). Independent tuning followed by fresh calibration also retains marginal coverage (Wu et al., 2026). Training-conditional coverage (Vovk, 2012; Bian & Barber, 2023) separates validity conditional on one calibration sample from validity on average over samples, and controls coverage for most of them. Our fit-conditional target makes the same separation for the interval score: it averages over a new size-m pool instead of certifying coverage. Cross-validation distinguishes risk estimation from model selection (Arlot & Lerasle, 2016; Wager, 2020; Bates et al., 2024). Our construction isolates calibration randomness within that distinction. Evaluating a model together with its post-hoc calibration, as calibrated log-likelihood does for classifiers (Ashukha et al., 2020), and averaging split-conformal set size over the calibration draw (Dhillon et al., 2024) are the closest precedents; we make the fresh-pool average exact for the interval score.

Computation and concentration. Complete U-statistics and conditional averaging are classical. Exact cross-validation formulas and incomplete-U optimization separate statistical targets from computational cost (Celisse & Robin, 2008; Clémençon et al., 2016). General exact-U algorithms exploit multiplicatively decomposable kernels (Chen et al., 2026); our kernel belongs to this class. We exploit its ordered comparisons and two hinges to compute the full risk and all fixed-m deletions in quasi-linear time (Appendix A.1). Approximate leave-one-out conformal methods accelerate predictor refits (Cong & Liu, 2026); our emitters remain fixed. Empirical Bernstein bounds for U-statistics already use kernel variance (Peel et al., 2010); projection variance and kernel interactions yield more refined complete- and incomplete-U bounds (Maurer, 2022), including growing-order

processes (Minsker & Yao, 2026). Bounded contrasts of unbounded scores likewise have precedents in sequential forecast comparison (Arnold et al., 2024, Appendix B). Our envelope uses the symmetrized calibration rank to remove the query-level $1/\alpha$ factor. The confidence rule applies established concentration and finite-family risk testing (Angelopoulos et al., 2025). Conformal risk control includes a U-risk extension (Angelopoulos et al., 2024); later stability-based work treats non-monotonic losses and multidimensional parameters (Angelopoulos, 2026). Our increment is the exact finite-rank calibrated score, its paired envelope and explicit information-cost examples. Bayesian quadrature (Snell & Griffiths, 2025) constructs an upper-risk random variable from scalar loss order statistics, a terminal loss bound and Dirichlet quantile spacings, unlike our query-excluded size-m subset average of calibration-dependent scores, so the two target different functionals.

Learning and adaptation. Conformal training, efficiency surrogates and boosting already optimize forecast quality with calibration in view (Stutz et al., 2022; Bellotti, 2020; Xie et al., 2024). Local regression, learned neighborhoods and distributional scores are established (Deshpande & Sarawagi, 2019; Woo et al., 2023; Lin et al., 2021; Guan, 2023; Chernozhukov et al., 2021). Our local fit is a matched intervention. Conformal regression efficiency already has asymptotic and finite-sample analyses (Sesia & Candès, 2020; Yao et al., 2026). SPCI, EnbPI, ACI and retrospective adjustment use sequential feedback (Xu & Xie, 2023; 2021; Gibbs & Candès, 2021; Jun & Ohn, 2026); their online feedback and validity conditions differ from independent final calibration, and we compare native versions in Appendix F. Classical predictive-accuracy tests compare forecast losses (Diebold & Mariano, 1995), conditional predictive-ability tests evaluate a forecasting method together with its finite estimation window (Giacomini & White, 2006), and sequential score comparisons extend this inference (Choe & Ramdas, 2024). Our target treats the finite calibration pool as part of the method in the same spirit, and we compute the complete-subset estimate of its expected interval score exactly.

7 Conclusion

Forecasts should be assessed under the calibration they will receive. The rank-and-hinge algorithm makes the complete finite-pool estimate and every fixed-m deletion exactly computable in quasi-linear time, with measured savings against a batched same-output reference. The sharp paired envelope makes controlled comparison possible without bounding the response, and the label-cost bounds show what reliable selection costs beyond exact assessment. A practitioner can therefore report, for each candidate, the exact fresh-calibration risk estimate at the deployment pool size together with its delete-one sensitivity, and act on a difference through Algorithm 1 when certified endpoint bounds are available. Together these results give a reproducible way to study which forecast differences survive fresh calibration, and when the available evidence is sufficient to act on them.

8 Limitations

The paired envelope and the controlled comparison require globally nearby endpoints, a common bounded scale, finite risks and IID selection labels independent of fitting. Observed sample maxima and chronological splits do not establish these conditions, and the effective radius also needs signed correction, because nonnegative correction does not inherit the invariance. The guarantee concerns the fresh mean risk, not every realized final pool or dependent rows, and the paper gives no estimable certificate for an unknown population. The constant 8 is uniform across ranks, not optimal at each rank. The label bounds are classical two-point constructions on atomic families; the fixed-level bound needs sufficiently large pools, and neither gives smooth-population or general minimax rates. The sufficient power calculation is not a required budget, and no forecasting comparison here establishes its margin under the global geometry. Exact computation leaves the statistic and its minimizer unchanged, and lower estimator variance can even raise selected risk, from $143/64$ to $163/64$ in an exact example under a declared tie rule (Theorem 3). The fixed-pool construction shows a failure of fixed-pool assessment, not harm from every added query, and a stronger fitted candidate, such as our quantile-regression controls, can remove it. Paired jackknife values for all A^2 pairs need additional covariance work, and near-tie floating-point decisions need error control beyond the real-arithmetic formula. Variance control under residual regularity does not by itself establish useful selection: its local examples remain weaker than ordinary range bounds, and in a constructed known-structure distribution classical DKW gives a tighter radius than paired variance, $.002363$ versus $.012078$, and

analytically permits a beneficial update with probability at least .95. This is an analytic conclusion for that artificial class, not a measured fitted-model gain (Appendix N.9).

The timing results are computational savings against our own dense reference, not a test of a general exact-U framework's software, and they strengthen neither confidence nor prediction. The unbounded-support study uses fixed analytic candidates and establishes no trained-model gain. With 32 repetitions per setting, zero observed harmful updates give only an 8.94 percent one-sided 95 percent upper bound on a setting's harm rate and do not establish simultaneous control; 20 versus 18 favorable settings do not establish a significant difference. When a tighter support bound is known, it should be used: with known binary support, an ordinary range envelope of 2 first permits updates, among the tested budgets, at 40, 80 and 160 labels ($\alpha = .20, .10, .05$ in the deterministic example), whereas the earlier rank bound with constant 12 first does so at 2,560, 5,120 and 10,240 (Appendix B). In a separate fitted comparison with 400 fit labels per repetition, all five controlled rules make zero updates across 256 fits, even at $N/(m+1) = 512$; in a post hoc diagnosis, the largest estimated benefit is 0.711 of its Bernstein radius. Raw-score minimization updates in 139 fits, 28 of them worse on independent final pools. The bounded candidate fits its coefficient without residualizing the feature against the linear baseline, which strongly attenuates it (median ratio .0733), so the bounded-versus-joint comparison partly reflects this estimator construction (Appendix B.5).

The forecasting rows are dependent, so the IID theorems are not invoked for them. The nine datasets informed development, and the matched result uses a Student-t decoder after a Gaussian version failed its prespecified criteria (Appendix E). Attenuation is not guaranteed. On seven subhourly datasets the control remains better, with retention $.7414$ $[.2335, 1.1114]$ ($R = +.03800$, $F = +.02817$), and a post hoc omission of the single-series Finland Traffic dataset reverses both contrasts ($R = -.01973$, $F = -.01784$). A preregistered daily replication was withdrawn after an input-normalization defect; the corrected secondary daily comparison retains $.9978$ of its contrast, or 1.0115 under equal-row weighting. On two chronological sources with assumed release delays, every tested pre-calibration selector has a higher aggregate final score than each of three fixed quantile-regression controls; in each setting an online rule that receives additional mature feedback scores lowest, while EnbPI, also online, scores worse than every control (Appendices F and F.2). No uniformly superior predictor or label-allocation rule follows.

AI USE STATEMENT

Large language model assistants were used as research and writing tools: for literature search, discussing ideas, writing code, checking derivations and editing the text. The authors directed the work, verified every proof, result, number and citation, and take full responsibility for the paper and its supplementary material.

ETHICS STATEMENT

The study evaluates uncertainty procedures on public time-series benchmarks; one appendix stress test also uses a proprietary daily record of fundraising campaigns provided by an industry partner and used for research only, in which campaigns are organizations rather than individuals. The study does not involve human-subject intervention. Each formal guarantee is limited to its declared data, candidates, losses, assumptions and selection rule. It must not be interpreted as a safety guarantee outside that scope. We report unfavorable and unresolved results alongside favorable ones because hiding them can create unjustified confidence in consequential forecasting systems.

REPRODUCIBILITY STATEMENT

Appendices A and B give the exact subset-risk computation, the paired envelope, the label-cost proofs and the full synthetic protocol. Appendix C gives the local-fit assumptions and proofs. Appendix E then records chronological roles, sample identity, metrics, complete result surfaces, and claim boundaries. The companion package reconstructs the complete finite-population and native online-comparison tables from exact expectations and emitted intervals, and the external-test estimates and intervals from saved series-level scores. It includes source code, protocol details, hashes, exclusions and data attribution. It does not reproduce every historical fit or redistribute unlicensed native software. Comparator implementations and information rights are stated separately.

REFERENCES

Karim Abou-Moustafa and Csaba Szepesvári. An exponential Efron–Stein inequality for L_q stable learning rules. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pp. 31–63, 2019. URL https://proceedings.mlr.press/v98/abou-moustafa19a.html.

Anastasios N. Angelopoulos. Conformal risk control for non-monotonic losses, 2026. URL https://arxiv.org/abs/2602.20151.

Anastasios N. Angelopoulos, Emmanuel J. Candès, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Sylvain Arlot and Matthieu Lerasle. Choice of V for V-fold cross-validation in least-squares density estimation. *Journal of Machine Learning Research*, 17(208):1–50, 2016. URL https://www.jmlr.org/papers/v17/14-296.html.

Sebastian Arnold, Georgios Gavrilopoulos, Benedikt Schulz, and Johanna Ziegel. Sequential model confidence sets, 2024. URL https://arxiv.org/abs/2404.18678v4. Version 4, revised January 2026.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2020.

Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, 119(546):1434–1445, 2024. doi: 10.1080/01621459.2023.2197686.

Anthony Bellotti. Constructing normalized nonconformity measures based on maximizing predictive efficiency. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pp. 41–54, 2020. URL https://proceedings.mlr.press/v128/bellotti20a.html.

Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023. doi: 10.1214/23-EJS2145. URL https://arxiv.org/abs/2205.03647.

Stéphane Boucheron and Maud Thomas. Concentration inequalities for order statistics, 2012. URL https://arxiv.org/abs/1207.7209.

Luis Candanedo. Appliances energy prediction, 2017. URL https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction.

Alain Celisse and Stéphane Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics & Data Analysis*, 52(5):2350–2368, 2008. doi: 10.1016/j.csda.2007.10.002.

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza, Max Mergenthaler-Canseco, and Artur Dubrawski. N-HiTS: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Xingyu Chen, Lin Liu, and Ruiqi Zhang. On computing and the complexity of computing higher-order U-statistics, exactly. *Statistics and Computing*, 36(5):223, 2026. doi: 10.1007/s11222-026-10974-x.

Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010. doi: 10.3982/ECTA7880.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021. doi: 10.1073/pnas.2107794118.

Yo Joong Choe and Aaditya Ramdas. Comparing sequential forecasters. *Operations Research*, 72(4):1368–1387, 2024. doi: 10.1287/opre.2021.0792. URL https://arxiv.org/abs/2110.00115.

Stéphan Clémençon, Igor Colin, and Aurélien Bellet. Scaling-up empirical risk minimization: Optimization of incomplete U-statistics. *Journal of Machine Learning Research*, 17(76):1–36, 2016. URL https://jmlr.org/papers/v17/15-012.html.

Jiachen Cong and Jingbo Liu. Accelerating conformal prediction via approximate leave-one-out. *arXiv preprint arXiv:2606.31915v1*, 2026. URL https://arxiv.org/abs/2606.31915v1.

Sayan Das, Bahram Yaghooti, Todd A. Kuffner, and Soumendra N. Lahiri. On optimal data splitting for split conformal prediction, 2026. URL https://arxiv.org/abs/2606.31600v1.

Prathamesh Deshpande and Sunita Sarawagi. Streaming adaptation of deep forecasting models using adaptive recurrent units. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019. doi: 10.1145/3292500.3330996. URL https://arxiv.org/abs/1906.09926.

Guneet S. Dhillon, George Deligiannidis, and Tom Rainforth. On the expected size of conformal prediction sets. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, 2024.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. doi: 10.1080/07350015.1995.10524599.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 1981. doi: 10.1214/aos/1176345462.

Shai Feldman, Stephen Bates, and Yaniv Romano. Improving conditional coverage via orthogonal quantile regression. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.00394.

Raffaella Giacomini and Halbert White. Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578, 2006. doi: 10.1111/j.1468-0262.2006.00718.x.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672, 2021.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Shelby J. Haberman. Cross-validation and U-statistics. *ETS Research Report Series*, 2019. doi: 10.1002/ets2.12263. URL https://doi.org/10.1002/ets2.12263.

Jonas Hanselle, Alireza Javanmardi, Tobias Florin Oberkofler, Yusuf Sale, and Eyke Hüllermeier. Conformal prediction without nonconformity scores. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1626–1639, 2025. URL https://proceedings.mlr.press/v286/hanselle25a.html.

Xuming He, Kean Ming Tan, and Wen-Xin Zhou. Robust estimation and inference for expected shortfall regression with many regressors, 2022. URL https://arxiv.org/abs/2212.05565v1.

Mahmoud Hegazy, Liviu Aolaritei, Michael I. Jordan, and Aymeric Dieuleveut. Valid selection among conformal sets. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2506.20173.

Nong Minh Hieu and Antoine Ledent. On finite-sample concentration of median of incomplete U-statistics, 2026. URL https://arxiv.org/abs/2606.00661v3.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

John Hogue. Metro interstate traffic volume, 2019. URL https://archive.ics.uci.edu/dataset/492/metro+interstate+traffic+volume.

Emilien Joly and Gábor Lugosi. Robust estimation of U-statistics. *Stochastic Processes and their Applications*, 126:3760–3773, 2016. URL https://arxiv.org/abs/1504.04580.

Jungbin Jun and Ilsang Ohn. Online conformal inference with retrospective adjustment for faster adaptation to distribution shift. *Pattern Recognition*, 180:114406, 2026. doi: 10.1016/j.patcog.2026.114406.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after data-dependent model selection. *Journal of the American Statistical Association*, pp. 1–13, 2026. doi: 10.1080/01621459.2026.2663588. Published online 29 June 2026.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Locally valid and discriminative prediction intervals for deep learning models. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/file/46c7cb50b373877fb2f8d5c4517bb969-Paper.pdf.

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*, 2025.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Andreas Maurer. Exponential finite sample bounds for incomplete U-statistics, 2022. URL https://arxiv.org/abs/2207.03136v1.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Andreas Maurer and Massimiliano Pontil. Empirical bounds for functions with weak interactions. In *Proceedings of the 31st Conference on Learning Theory*, volume 75, pp. 987–1010. PMLR, 2018. URL https://proceedings.mlr.press/v75/maurer18a.html.

Stanislav Minsker. U-statistics of growing order and sub-Gaussian mean estimators with sharp constants, 2023. URL https://arxiv.org/abs/2202.11842v3.

Stanislav Minsker and Shunan Yao. Deviation inequalities for U-processes of growing order, with applications to half-space depth, 2026. URL https://arxiv.org/abs/2608.12056v1.

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.

Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. Empirical Bernstein inequalities for U-statistics. In *Advances in Neural Information Processing Systems*, volume 23, 2010. URL https://papers.neurips.cc/paper/4081-empirical-bernstein-inequalities-for-u-statistics.

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's TIME: Towards the next generation of time series forecasting benchmarks. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.12147.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, 2019.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020.

Matteo Sesia and Emmanuel J. Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020. doi: 10.1002/sta4.261. URL https://arxiv.org/abs/1909.05433.

Jake C. Snell and Thomas L. Griffiths. Conformal prediction as Bayesian quadrature. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56068–56084, 2025. URL https://proceedings.mlr.press/v267/snell25a.html.

David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.09192.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, 2012.

Stefan Wager. Cross-validation, risk estimation, and model selection: Comment on a paper by Rosset and Tibshirani. *Journal of the American Statistical Association*, 115(529):157–160, 2020. doi: 10.1080/01621459.2020.1727235.

Yuhao Wang and Tengyao Wang. Localized conformal model selection. *arXiv preprint arXiv:2602.19284*, 2026. URL https://arxiv.org/abs/2602.19284.

Fan Wei and Richard M. Dudley. Dvoretzky–Kiefer–Wolfowitz inequalities for the two-sample case, 2011. URL https://arxiv.org/abs/1107.5356v2. Version 2.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Learning deep time-index models for time series forecasting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37217–37237. PMLR, 2023. URL https://proceedings.mlr.press/v202/woo23b.html.

Fanyi Wu, Lihua Niu, Samuel Kaski, and Michele Caprio. Decoupled conformal optimisation: Efficient prediction sets via independent tuning and calibration, 2026. URL https://arxiv.org/abs/2605.18354v1.

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*, 2021.

Junxi Wu, Dongjian Hu, Yajie Bao, Shu-Tao Xia, and Changliang Zou. Error-quantified conformal inference for time series. In *International Conference on Learning Representations*, 2025.

Ran Xie, Rina Foygel Barber, and Emmanuel J. Candès. Boosted conformal prediction intervals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 71868–71899, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/842714f78c95096e20ac7d2591c5a24b-Abstract-Conference.html.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2021.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, 2023.

Yachong Yang and Arun Kumar Kuchibhotla. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, 120(549):435–447, 2025. doi: 10.1080/01621459.2024.2344700. URL https://arxiv.org/abs/2104.13871.

Yunzhen Yao, Lie He, and Michael Gastpar. Non-asymptotic analysis of efficiency in conformalized regression. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/467b55d1eeee930e0316d539ca5abd3d-Paper-Conference.pdf.

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*, 2021.

Zheng Zhou, Xiangfei Zhang, Chongguang Tao, and Yuhong Yang. Conformal prediction assessment: A framework for conditional coverage evaluation and selection. *arXiv preprint arXiv:2603.27189*, 2026. URL https://arxiv.org/abs/2603.27189.

GUIDE TO THE SUPPLEMENTARY MATERIAL

The exact computation and paired-score proofs come first (Appendices A–B). Appendix C gives the local-fit assumptions and proofs, Appendix D the full experimental results and comparator fidelity, and Appendix E the reproducibility and evaluation scope. Appendix F reports the chronological native comparisons, Appendix G the finite-calibration averaging results, and Appendix H the fixed-pool construction. Appendices I–K collect the earlier endpoint, cancellation and selection analyses, Appendix L the retained forecasting protocols, and Appendix M the complete forecasting tables, including the separately specified attenuation test (Appendix M.6). The fixed-level label bounds and regularity conditions close the appendix (Appendix N). Each study keeps its own statistical units and limitations.

A EXACT COMPUTATION OF THE COMPLETE-SUBSET SCORE

The ordinary complete-subset control need not enumerate calibration subsets. For a fixed forecast, its interval score is piecewise linear in the calibration correction. This permits an exact weighted-prefix calculation after one sort. The result changes computational cost, not the statistical estimator or its selection risk. Exact cross-validation formulas have precedents in other losses and models (Celisse & Robin, 2008); incomplete U-statistic optimization also explicitly separates sampling from computational approximation (Clémençon et al., 2016).

Setting. Condition on ordered endpoints (L_j, U_j), positive scales c_j, and outcomes Y_j, for $j = 1, \dots, N$. A single scale applies to both tails. Let $N \geq m + 1$, $0 < \alpha < 1$, and $k = \lceil (m+1)(1-\alpha) \rceil \leq m$. For each query j, average the issued interval score over all size-m calibration subsets of the other $N - 1$ observations. The query is never in its own calibration subset. Signed joint residuals and their nonnegative truncation are both supported; crossed endpoints collapse to their original midpoint. These identities hold for any fixed array. Interpreting their expectation as fresh-calibration risk additionally requires the independent sampling and fixed-candidate assumptions used above.

Subset probabilities. Sort $r_i = \max(L_i - Y_i, Y_i - U_i)/c_i$, breaking ties by observation index, and let h_j be query j's one-based position. Write $n = N - 1$. The probability that rank r is the kth member of a uniformly chosen size-m subset is

$$w_r = \frac{\binom{r-1}{k-1}\binom{n-r}{m-k}}{\binom{n}{m}}, \qquad r = 1, \dots, n. \tag{6}$$

Invalid binomial coefficients are zero. The numerator chooses the required indices on either side of rank r, so the weights sum to one. In the global sort, position $i < h_j$ receives weight w_i, position $i > h_j$ receives w_{i-1}, and the query receives zero. This counts subsets of indices and remains valid with ties. Let $t_i = r_i$ for signed calibration and $t_i = \max(r_i, 0)$ otherwise. Truncation preserves order.

Score identity. For a query write $z = (L + U)/2$, $v = (U - L)/2$, $d = |Y - z|$, $a = -v/c$, and $b = (d - v)/c$. Its issued half-width is $\max(v + ct, 0)$, and

$$\ell(t) = \frac{2d}{\alpha} + \left(2 - \frac{2}{\alpha}\right) c(t - a)_+ + \frac{2c}{\alpha}(t - b)_+. \tag{7}$$

Below a the interval collapses and the loss is $2d/\alpha$. Between a and b the outcome lies outside the interval, giving slope $(2 - 2/\alpha)c$. Above b it is inside, giving slope $2c$. Both junctions are continuous, including $d = 0$. This proves the identity on all segments.

Prefix evaluation. Build cumulative weight and weighted-t sums for the two arrays $(w_1, \dots, w_n, 0)$ and $(0, w_1, \dots, w_n)$. On any contiguous range, $\sum_i w_i(t_i - x)_+$ equals the weighted-t sum above x minus x times the corresponding weight sum. Binary search finds the first value above x; two prefix differences give each sum. Apply the first array below h_j and the second above it, at both $x = a$ and $x = b$. Equation 7 then gives the exact query-excluded score. Average over j.

Each candidate requires $O(N \log N)$ arithmetic operations and $O(N)$ workspace. The weights can be generated in linear unit-cost arithmetic from the first positive weight using

$$\frac{w_{r+1}}{w_r} = \frac{r}{r - k + 1} \frac{n - r - m + k}{n - r}$$

on their positive support. This is not a bit-complexity claim for exact rationals. For A fixed candidates, the cost is $O(AN \log N)$; retaining all per-query scores uses $O(AN)$ storage. Exact arithmetic gives the same choice as ordinary complete-subset minimization under the same tie rule. It does not reduce the sampling error of that estimator. If $k > m$, the finite formula is inapplicable and the declared score is infinite.

Implementation check. A frozen synthetic matrix contains 896 cases covering $N = 2, \dots, 7$, every finite rank, signed and nonnegative corrections, ties, zero widths and unequal positive scales. Literal

subset enumeration reproduces all 5,152 per-query rational scores exactly. A separate process repeats this enumeration without importing the prefix implementation. This is an independent computation, not an independent authorship or proof review. Floating-point prefix results differ from these exact scores by at most 6.40×10^{-14}. Thirteen invalid-input checks and two overflow checks also pass. Finite checks do not prove the identity or certify floating-point accuracy for arbitrary ranges or near ties.

Table 4: Computing the same ordinary score on supplied synthetic arrays. For each N, ranges summarize six cells: $m = 1, 9, 19$ at $\alpha = 1/2, 1/10, 1/20$, respectively, with signed and nonnegative corrections. Each cell uses three candidates and the median of three alternating-order timings on one CPU. The ratio is per-query-sort time divided by prefix time. These are implementation measurements, not uncertainty intervals or selection performance comparisons. All 24 cells and all calls are in the companion.

N	Prefix (ms)	Per-query sort (ms)	Time ratio
32	0.67–0.85	3.56–3.82	4.32–5.41
128	1.01–1.13	19.08–19.59	17.27–18.90
512	1.36–1.65	112.31–115.31	68.48–85.08
2048	2.33–3.07	760.90–1026.47	319.12–413.45

Each measured call includes rank weights, residuals, sorting and averaging; input generation, fitting, final calibration and testing are excluded. The reference implementation sorts separately for every query. Another optimized implementation of the ordinary criterion may be equally fast or faster. Floating-point objective differences between the two implementations are at most 7.00×10^{-13} in this timing matrix. Prefix subtraction can suffer cancellation; the routine does not round ties or clip negative rounding errors. No historical forecasting experiment is rerun, no general allocation rule is validated, and no first-in-literature claim follows from this computation.

Single-sort mean-only reference. A second implementation sorts once and vectorizes the query-excluded evaluation. It returns the same complete mean in $O(AN^2)$ work. For three candidates, $m = 9, 19$, and both calibration operators, the medians over three alternating calls are below. These outputs differ from the two-candidate all-deletion comparison in Table 2.

Table 5: Mean-only evaluation against a single-sort reference. Ranges span four settings per N; fitting and final evaluation are excluded. The settings differ from those of Table 4, so absolute times are comparable only within each table.

N	Prefix (ms)	Single-sort reference (ms)
128	0.62–0.68	1.42–1.50
512	0.85–0.93	13.77–14.48
2048	1.72–2.02	220.89–243.87

A.1 All delete-one risks at the same cost

Removing one observation removes both its query and calibration roles. Keep m and k fixed and require $N \geq m + 2$. In the residual ordering above, write $\ell_j(t)$ for Equation 7, $A_r = \sum_{j<r} \ell_j(t_r)$, and $B_r = \sum_{j>r} \ell_j(t_r)$. In this subsection the weights are

$$w_r = \frac{\binom{r-1}{k-1}\binom{N-2-r}{m-k}}{\binom{N-2}{m}}, \qquad w_r = 0 \text{ outside } 1 \leq r \leq N-2.$$

Let U_{-i} denote the complete score after deleting row i. Counting the position of the calibration threshold after excluding both i and the query gives

$$(N-1)U_{-i} = \sum_{r<i}(w_r B_r + w_{r-1}A_r) + \sum_{r>i}(w_{r-1}B_r + w_{r-2}A_r) - \sum_{r<i} w_r \ell_i(t_r) - \sum_{r>i} w_{r-2}\ell_i(t_r). \tag{8}$$

The two subtractions remove the deleted query; the threshold $r = i$ never appears. The denominator uses a calibration universe of size $N - 2$, not a changed calibration size. This proves the formula including ties.

For $j < r$, the query is covered at t_r, so $A_r = 2\sum_{j<r} v_j + 2t_r \sum_{j<r} c_j$. Sorting all hinge breakpoints gives the all-query sum $H(t_r) = \sum_j \ell_j(t_r)$ by prefix evaluation. Then $B_r = H(t_r) - A_r - \ell_r(t_r)$. Ordinary prefix and suffix sums compute the first line of equation 8; weighted hinge sums compute the second. All N deletions therefore cost $O(N \log N)$ arithmetic and $O(N)$ workspace per candidate. An alternative Fenwick implementation gives the same values. Restore original row order before pairing different residual sorts. The API requires matching observation identities, α, m, k and operator.

Every size-d subset appears in $N - d$ deletion datasets, hence $U_N = N^{-1}\sum_i U_{-i}$. For a predetermined pair, apply this identity to the risk difference and compute

$$V_J = \frac{N-1}{N}\sum_i (U_{-i} - U_N)^2.$$

The $O(AN \log N)$ cost covers A candidates and $O(A)$ specified comparisons, such as comparison to one reference. Arbitrary all-pairs jackknife variances require additional covariance computation, directly $O(A^2 N)$.

Independent computation and timing. Literal subset enumeration checks 306 rational cases and 7,524 risk values, including $m = 1$, $N = m + 2$, ties, collapse and both operators. Seven invalid-input checks include equal-rank but different-α pairs and reversed row identities. For two supplied candidates at $N = 512$, the four $m \in \{9, 19\}$ and operator settings take 9.3–9.5 milliseconds, versus 508–528 milliseconds for N calls to the existing complete-score routine. At $N = 32{,}768$, the fast calculation takes 0.78–0.80 seconds; no repeated-complete timing is reported at that size. Calls include weights and sorting, not data generation or fitting. Fast timings are medians of three calls and the repeated-complete reference has one call, in fixed order. These describe implementations, not sampling uncertainty or selection benefit. All 20 timing cells are retained in the companion.

A batched dense comparison. A stronger reference need not repeat the full calculation N times. Materialize $L_{jr} = \ell_j(t_r)$, form its triangular column sums A_r, B_r, and evaluate Equation 8 with prefix sums and row reductions. This costs $O(N^2)$ arithmetic and storage for all deletions together. Our vectorized implementation fills the matrix in 256-row blocks, using literal issued endpoints and explicit midpoint collapse. It is our structured dense reference, not the software of Chen et al. (2026).

Table 6 reports every setting of a separate same-array comparison. Each call returns two full risks, both complete deletion vectors and their paired jackknife. All 306 rational cases pass literal-subset checks; every floating difference over the twelve settings is at most 3.26×10^{-13}. Only after every vector passed were timings measured. Prefix calculation is faster in these settings, by 1.44–1.46 times at $N = 512$ and 24.65–24.69 times at $N = 8192$. These are descriptive implementation ratios, not evidence of stronger inference or dominance over other implementations.

Methods run in alternating order with one numeric-library thread. Each fresh worker performs one untimed warm-up and one measured call; weights, sorting, allocation and pairing are included. Input generation, imports and startup are excluded, with startup-inclusive times retained separately. Memory uses the current process image's Linux high-water mark in a separate untimed pass and includes interpreter/library overhead. The original resource-usage peaks inherited a parent high-water mark and are retained as invalid measurements; no timings were repeated. The companion retains all 72 calls and input hashes.

Table 6: All-deletion computation: median time over three calls and process peak memory. Ratio is dense time divided by prefix time.

			Time (ms)			Memory (MiB)	
N	m	Correction	Prefix	Dense	Ratio	Prefix	Dense
512	9	Signed	9.65	13.99	1.45	47.1	54.8
512	9	Nonnegative	8.50	12.33	1.45	51.2	54.6
512	19	Signed	8.39	12.24	1.46	47.2	54.5
512	19	Nonnegative	8.50	12.24	1.44	49.1	54.8
2,048	9	Signed	34.52	213.57	6.19	49.2	99.6
2,048	9	Nonnegative	34.64	210.97	6.09	48.9	99.6
2,048	19	Signed	34.56	209.60	6.07	49.2	99.7
2,048	19	Nonnegative	34.43	210.45	6.11	49.0	99.6
8,192	9	Signed	144.69	3566.63	24.65	56.6	658.0
8,192	9	Nonnegative	144.52	3568.16	24.69	58.2	658.0
8,192	19	Signed	144.74	3572.77	24.68	56.5	657.9
8,192	19	Nonnegative	144.92	3576.21	24.68	56.6	658.0

Relation to general exact computation. Let ρ order residuals, breaking ties by supplied row position. For an ordered distinct $(m+1)$-tuple, take the query at position zero, the threshold at position one, and define the star-product kernel

$$F(z_0, \dots, z_m) = \ell_0(t_1) \prod_{s=2}^{k} \mathbf{1}\{\rho_s < \rho_1\} \prod_{s=k+1}^{m} \mathbf{1}\{\rho_s > \rho_1\}.$$

Each query and calibration subset contributes $(k-1)!(m-k)!$ equal terms. Thus $m\binom{m-1}{k-1}$ times the distinct-tuple sum, divided by $(N)_{m+1}$, equals U_N. This places the statistic in the decomposable framework of Chen et al. (2026). Our additional calculation exploits ordered comparisons and two hinges, rather than claiming a new general exact-U framework. Generic inclusion sums can also recover all deletions: for $W(x) = \sum_{|S|=d} h(S) \prod_{i \in S} x_i$, $U_{-i} = [W(\mathbf{1}) - \partial_i W(\mathbf{1})]/\binom{N-1}{d}$. Efficient differentiation need not make N independent calls. Approximate conformal refitting (Cong & Liu, 2026) addresses a different object: here the predictor is fixed and only the finite calibration/query pool changes.

A.2 Classical inference from the deletion values

Condition on independently fitted candidates and IID observations from the fresh-risk population. For the fixed paired kernel $h = K_a - K_b$ of degree $d = m+1$, assume finite individual risks and a known global $|h| \le B$. Let its canonical projection variances be ζ_s. Orthogonality gives

$$\mathbb{E}V_J = \sum_{s=1}^{d} \frac{s(N-1)}{N-s} \frac{\binom{d}{s}^2 \zeta_s}{\binom{N}{s}}.$$

The first-order multiplier is one, while higher-order multipliers exceed one. This classical expectation conservatism (Efron & Stein, 1981) is not a samplewise or high-probability upper bound. For example, with $m = 1$, $\alpha = 1/2$, unit scale, zero raw widths, centers $0, a$ with $0 < a < 1/2$, and $Y \sim \mathrm{Bernoulli}(1/2)$, the paired kernel at outcomes $00, 01, 11$ is $2a, -4a, -2a$. Its expectation is $-2a$; the all-zero sample has $U_N = 2a$ and $V_J = 0$ with probability 2^{-N}.

A valid full-N bound follows from classical results, not from dropping the higher orders. Put $n = N-1$, $x = \log(2/\delta)$ and

$$V = \sum_i (U_{-i} - U_N)^2, \qquad M_t = \frac{2Bd}{t}, \qquad J_t = \frac{4Bd(d-1)}{t-1}.$$

For $N \ge d+1$, a one-sided radius about U_N is

$$r_N = \sqrt{2x}\left\{\sqrt{V} + \sqrt{(2M_n^2 + 8J_n^2)x}\right\} + (2M_N/3 + J_N)x. \tag{9}$$

Indeed, the expected Efron–Stein variance of the same kernel on t rows is $S_t = \sum_{s=1}^{d} s\binom{d}{s}^2 \zeta_s / \binom{t}{s}$, so $S_N \le S_n$. The variance functional of Maurer & Pontil (2018, Eq. 10) for U_n on $n+1=N$ observations equals V. Their one-sided Theorem 2 bounds $\sqrt{S_n}$ by $\sqrt{V} + \sqrt{(2M_n^2 + 8J_n^2)x}$ with failure probability $\delta/2$. Their Eq. 9 applied to $-U_N$ gives $\mathbb{E}h - U_N \le \sqrt{2S_N x} + (2M_N/3 + J_N)x$ with the same failure allocation. A union bound proves equation 9, retaining all orders. Multiple comparisons require a declared error allocation. This is not an anytime statement or a new concentration inequality.

Geometry and a matched ordinary comparison. For the timing inputs $X \in [-1,1]$, $c(X) = 1 + X/4$, centers $0.4X$ and $0.4X + 0.1\sin(3X)$, and shared half-width $c(X)$, the raw endpoint gap is at most $1/10$, but the standardized gap requires $\epsilon = 2/15$ since $c \ge 3/4$. Using $c_{\max} = 5/4$ and $\alpha = 1/10$, the general bound is $B = 4/3$. The first calculation incorrectly used $1/10$ as the standardized gap; the companion preserves it and the separately checked confidence-only correction. No fitting or timing was repeated. The existing center-only bound also proves $B = 1$ for this pair, and Equation 12 gives $14/15$. These distinct justifications do not validate the original invocation. The following comparison retains the conservative corrected $4/3$.

At $N = 32{,}768$, signed correction and $\delta = .05$, the full-N radii are .277339 and 1.150705 for $m = 9$ and 19, respectively. Ordinary U-Bernstein gives .016167 and .023838 using the same labels and global range. Across all 20 cells, the classical full-N radius is never smaller. The generic mixed-interaction cost is of order Bd^2/N. Fast deletion is a computational benefit, not evidence of a better confidence rule or selector.

A.3 Why endpoint closeness alone cannot remove the interaction cost

For the two calibration sizes above, the global interaction order cannot be reduced using endpoint geometry alone. Choose unit scale, common raw half-width $w > 0$, centers $0, a$ with $a > 0$, and outcomes $-L, M$ with $L > w$ and $M > L + 2a$. Both operators coincide because all residuals are positive. Write $q = d - k$. For a block with r high outcomes, direct query exclusion gives

$$h(r) = \begin{cases} 2a - 4ar/(\alpha d), & r < q, \\ -2a + (4aq/d)(1 - 1/\alpha), & r = q, \\ -2a, & r > q. \end{cases}$$

At $\alpha = .1$ and $d = 10, 20$, the exact envelope is $B = 28a/5$. The second differences $h(2) - 2h(1) + h(0)$ are $2B$ and $-9B/14$. For any $N \ge d + 1$, let all rows except two replaceable observations be low. Only subsets containing both replaced rows survive the mixed difference. Their fraction is $d(d-1)/(N(N-1))$. Thus the global seminorm $J(U_N) = N \sup |D_i D_j U_N|$ satisfies

$$J(U_N) \ge \begin{cases} 2Bd(d-1)/(N-1), & d = 10, \\ (9/14)Bd(d-1)/(N-1), & d = 20. \end{cases}$$

These are respectively $1/2$ and $9/56$ of the generic upper bound, for every admissible N. Shrinking a shrinks both bounds together. This is a worst-case obstruction for these ranks, not a claim about its frequency or all fixed-α, growing-d regimes. An argument based on likely rather than global interactions needs additional justified assumptions. The companion checks 64 literal interval kernels, 48 ordered replacement configurations and 12 fresh-query population identities exactly; finite checks support the implementation, while the all-N conclusion follows from counting.

B A PAIRED SCORE BOUND WITHOUT BOUNDED RESPONSES

An absolute interval score may be arbitrarily large. A comparison between two nearby forecasts need not be. The finite calibration rank provides a further reduction when every point in a small block is evaluated once as the query. This gives a usable range for ordinary confidence-based selection; it does not introduce a new model or testing principle.

Setting. Condition on two independently fitted, ordered endpoint functions, indexed by a, b. They share the same positive scale $c(x) \le c_{\max}$ for both tails. Suppose the following bound holds on the entire input support:

$$\max\{|L_a(x) - L_b(x)|, |U_a(x) - U_b(x)|\} \le \epsilon c(x). \tag{10}$$

The constants must be known from the fitted functions and their domain, not estimated by a maximum over observed queries. Set $d = m + 1$, $k = \lceil d(1-\alpha) \rceil \le m$, and $s = d - k \le \alpha d$. For each d-point block, let K_a average the d interval scores obtained by using each point as a query and the other m points as calibration. Use the joint residual, signed or nonnegative correction, and midpoint collapse consistently across candidates. Assume the reference's fresh-calibration risk is finite.

Proof of Proposition 2. The bound is

$$B_{ab} = c_{\max}\epsilon \min\{4/\alpha,\ 4 + 4s/(\alpha d)\} \le 8c_{\max}\epsilon.$$

The earlier union-of-missed-sets argument gave the valid but looser constant 12. Its frozen numerical comparison below is retained unchanged; it does not use the tightened constant.

Proof. The two standardized residual vectors differ by at most ϵ in sup norm. The order statistic and nonnegative truncation are both 1-Lipschitz, so each query's correction differs by at most ϵ. Its endpoints before collapse differ by at most $2c(x)\epsilon$. Midpoint collapse is the projection onto ordered endpoint pairs and is 1-Lipschitz in sup norm. Therefore, the two issued widths differ by at most $4c(x)\epsilon$, and their distances from the outcome differ by at most $2c(x)\epsilon$.

A missed query has residual strictly above its leave-one-out kth order statistic. Indeed, a residual no larger than the correction places the outcome between the uncollapsed endpoints, which must then be ordered. Nonnegative truncation can only enlarge them. Among d residuals at most $s = d - k$ satisfy the strict inequality; ties can only reduce that number. Write $h_j = \mathrm{dist}(Y_j, I_{a,-j}) - \mathrm{dist}(Y_j, I_{b,-j})$. A positive h_j can occur only where candidate a misses; a negative one only where b misses. Each sign has at most s contributing entries, of magnitude at most $2c_{\max}\epsilon$. Consequently $|\sum_j h_j| \le 2sc_{\max}\epsilon$, rather than the sum of absolute terms over the union of both missed sets. Averaging width and signed penalty differences gives $4c_{\max}\epsilon + 4sc_{\max}\epsilon/(\alpha d)$. Alternatively, the interval score is $2/\alpha$-Lipschitz in its ordered endpoints under the sup norm. Applying the endpoint bound separately to every query gives $4c_{\max}\epsilon/\alpha$. Take the smaller bound. □

A tighter bound for center-only changes. If the two raw intervals have the same half-width $v(x)$ at every input and their centers differ by at most $\epsilon c(x)$, their issued centers remain the raw centers and their issued half-widths are $[v(x) + c(x)q]_+$. The two half-widths differ by at most $c(x)\epsilon$, even if one interval collapses. Thus the width difference is at most $2c(x)\epsilon$, while the difference in distance from the outcome is at most $2c(x)\epsilon$. Applying the signed missed-query argument in the proof above gives the more specific bound

$$|K_a - K_b| \le c_{\max}\epsilon\left(2 + \frac{4s}{\alpha d}\right) \le 6c_{\max}\epsilon. \tag{11}$$

This applies to signed and nonnegative correction, including ties and midpoint collapse. It is not the bound for arbitrary endpoint changes in Proposition 2.

If the shared raw half-width is further $v(x) = \nu c(x)$ for one constant $\nu \ge 0$, the optimal constant within this subclass at each finite rank is

$$C_{d,k,\alpha} = 2 + \frac{4s}{\alpha d} - \frac{4s}{d}. \tag{12}$$

To see the upper bound, order the standardized absolute residuals $e_i = |Y_i - z(X_i)|/c_i$ and carry their scales with the ordering. Under signed correction, the raw ν cancels from each issued half-width. With $C_L = \sum_{j \le k} c_{(j)}$ and $C_H = \sum_{j>k} c_{(j)}$, the symmetric block score within a strict ordering region is

$$K(e) = \frac{2C_L}{d} e_{(k+1)} + \frac{2C_H}{d}(1 - 1/\alpha) e_{(k)} + \frac{2}{\alpha d} \sum_{j>k} c_{(j)} e_{(j)}.$$

Its gradient has ℓ_1 norm $2\bar{c}+4(1/\alpha-1)C_H/d \le c_{\max} C_{d,k,\alpha}$, where $\bar{c} = d^{-1} \sum_j c_j$. For nonnegative correction, replace each leave-one-out threshold by $\max(\nu, e_{(k)})$ or $\max(\nu, e_{(k+1)})$. If the threshold crosses ν, the gradient norm is at most $2C_L/d + 2C_H/(\alpha d)$; if both thresholds stay below ν, only residuals above ν contribute, with gradient norm at most $2C_H/(\alpha d)$. Both bounds are no larger than $c_{\max} C_{d,k,\alpha}$. The block score is continuous at residual ties and at ν, so the piecewise bound applies globally.

The constant is attained. Take $c = 1$, k common observations with outcome $\tau > \nu + \epsilon$, and s rare observations with outcome $M > \tau + 2\epsilon$. The reference center is zero; the other center shifts by $+\epsilon$ on common inputs and $-\epsilon$ on rare inputs. Both raw half-widths equal ν. Every correction is positive. For a common query the alternative's score rises by 2ϵ; for a rare query it rises by $4\epsilon/\alpha - 2\epsilon$. Their average equals $C_{d,k,\alpha}\epsilon$. This also approaches the uniform constant six along $\alpha = 1/d$ as d grows. These bounds sharpen a certified range, not the observed decisions of the completed fitted experiment.

The distinction between a query and a symmetrized block matters. With unit scale, constant point forecasts zero and ϵ, an all-zero calibration pool and query $Y = \epsilon$ give a score difference $(2/\alpha - 2)\epsilon$. A per-query range therefore can grow as $1/\alpha$. If every outcome is zero instead, the issued intervals are $[0, 0]$ and $[0, 2\epsilon]$. Their block difference is -2ϵ at every finite rank, so no uniform $o(\epsilon)$ envelope is possible.

A classical finite-label consequence. For $N \ge d$ IID observations from the same population defining R_m, independent of fitting, let $\widehat{D}_{ab}$ be the complete U-statistic with kernel $K_a - K_b$ and let $n = \lfloor N/d \rfloor$. Its mean is $D_{ab} = R_m(a) - R_m(b)$. The usual random-permutation block argument and Hoeffding bound yield

$$\Pr\{D_{ab} > \widehat{D}_{ab} + t\} \le \exp\{-nt^2/(2B_{ab}^2)\}.$$

For $A - 1$ candidates compared to a fixed reference, use $t_{ab} = B_{ab}\sqrt{2\log((A-1)/\delta)/n}$ and permit an update only if $\widehat{D}_{ab} + t_{ab} < 0$. Simultaneous control bounds the probability of a harmful update by δ. This is the standard risk-testing principle (Angelopoulos et al., 2025), not a new selection algorithm. For a genuinely beneficial candidate with gap $\gamma > t_{ab}$, the same bound gives a failure-to-certify probability at most $\exp\{-n(\gamma - t_{ab})^2/(2B_{ab}^2)\}$. The unknown gap is used only in this power statement. The rule does not need it. When $B_{ab} = 0$, keeping the reference suffices.

For two candidates and $\delta = \beta = .05$, a sufficient condition for certification with probability at least $1-\beta$ is $n > 8B_{ab}^2 \log(20)/\gamma^2$. Substituting the uniform bound $B_{ab} \le 8c_{\max}\epsilon$ yields the conservative condition $n > 512\log(20)(c_{\max}\epsilon/\gamma)^2$ used in the main text. This follows by applying the upper and lower concentration tails separately; it does not estimate γ from selection data or assert that the condition is necessary.

These are fresh-pool mean-risk statements, not guarantees for every realized final pool. They retain the sample factor $\lfloor N/(m+1) \rfloor$ and can remain conservative. Different candidate-specific scales, a sample-only geometry bound, dependent selection rows or infinite individual risks do not inherit this result. Numerical implementations also need valid roundoff control before the mathematical guarantee can be asserted for near-tie floating-point decisions.

A complete finite-law comparison. To isolate the role of the range, fix point candidates $a = 0, b = 1$, unit scale, $Y \sim \text{Bernoulli}(p)$ and $\alpha = 1/(m+1)$. The distribution parameter is unknown to every rule and used only for evaluation. There is no candidate training or pilot. With j ones in a d-point block and $m \ge 2$, the paired kernel equals -2, $2(d-1)/d$, $-2(d-1)/d$ and 2 at $j = 0, 1, d-1, d$, respectively, and zero otherwise; for $m = 1$ the two middle cases coincide, and the symmetry $j \mapsto d - j$ makes the kernel zero there. Consequently, the stronger ordinary control using the known binary support has the exact range 2. It receives the same support information as all other

rules and is retained rather than replaced by a loose baseline. Direct averaging over an independent final calibration pool gives

$$R_m(a) = 2\{1-(1-p)^m\} + \frac{2p}{\alpha}(1-p)^m, \qquad R_m(b) = 2(1-p^m) + \frac{2(1-p)}{\alpha}p^m.$$

For t ones among N selection labels, the complete-U difference is the expectation of the block kernel under $\mathrm{Hypergeometric}(N, t, d)$. Integrating each decision against $\mathrm{Binomial}(N, p)$ yields its selection risk without Monte Carlo error. This remains a floating-point evaluation of a finite sum, not an exact numerical certificate.

The fixed matrix uses $m \in \{4, 9, 19\}$, $N/d \in \{1, 8, 64, 512, 4096\}$ and $p \in \{0, .02, .1, .2, .3, .5, .7, .8, .9, .98, 1\}$, giving 165 distribution-and-budget cells. Six rules are evaluated in every cell: keep, raw-score minimization, complete-U minimization and three confidence rules using the per-query, rank-based or known-support range. Every tie favors keep. All use the same two candidates and final calibration size; all except keep consult the N selection labels. The 990 records, including harmful updates and inaction, are supplied in the companion. They are not independent datasets or forecasting replications.

Table 7: Smallest tested selection budget N that permits an update in the deterministic $p = 0$ example, with $\delta = .05$. The final-score improvement is exactly 2 in this example. "None" means none on the fixed grid, not an impossibility. The original 12-based rank bound improves on the per-query range, while the stronger ordinary support-aware bound needs fewer labels. This example illustrates nonvacuity, not superiority on an unknown collection of tasks.

α	Query bound	Rank bound	Known-support bound
0.20	20480	2560	40
0.10	40960	5120	80
0.05	none	10240	160

Exact rational checks cover 480 deterministic paired blocks, computing 11,808 issued query scores, including ties, negative corrections, collapse, variable shared scales and large outcomes. Six overflow configurations are rejected rather than clamped. A separate prefix computation gives identical block means; 38 binary-kernel and 33 population-risk identities also pass. These checks support the implementation, not the universal proof. Independent finite-sum recomputation of 2,970 saved quantities differs by at most 1.08×10^{-10}. It is a second computation by the same author, not an external review. Tiny nonzero harmful-update probabilities remain in the records; a computed zero may reflect floating-point underflow rather than an impossible event. The observed diagnostic does not establish a better learned model, a useful budget-allocation policy or performance on a new forecasting domain.

B.1 Sharpness, realizability and finite risk

For signed sharpness let $\epsilon > 0$, $K > 2\epsilon$, $M > K$, $\alpha = 1/d$ and $c = 1$. In a block of $d-1$ common points and one rare point, the common outcome is zero and the endpoints are $a = [-\epsilon, \epsilon]$, $b = [0, 0]$. The rare outcome is M and its endpoints are $a = [-K+\epsilon, K-\epsilon]$, $b = [-K, K]$. The endpoint geometry holds everywhere on this two-point input support. For a common query the two calibration maxima are $M - K + \epsilon$ and $M - K$, so the score difference is 4ϵ. For the rare query, calibration maxima are $-\epsilon$ and zero; the issued widths differ by -4ϵ and the outside distances by 2ϵ. Its score difference is $4\epsilon(d-1)$. The block mean is $8\epsilon(d-1)/d$, approaching 8 times ϵ. This proves uniform signed sharpness across ranks, not sharpness at each d. Nonnegative correction replaces the negative maximum by zero: the rare difference is $2\epsilon(d-1)$ and the block mean is $6\epsilon(d-1)/d$ in this particular example.

The same uniform constant is necessary for nonnegative correction. Keep the endpoints above but set every common outcome to $\tau > \epsilon$ and require $M > K + \tau$. Common residuals are $\tau - \epsilon, \tau$; rare residuals are $M - K + \epsilon, M - K$. All are strictly positive, and the rare residual exceeds the corresponding common one. At $\alpha = 1/d$, each leave-one-out correction is the largest residual. When a common point is the query, the rare point determines the correction; both issued intervals contain τ and their scores differ by 4ϵ. When the rare point is the query, the corrections are $\tau - \epsilon$ and τ. The

issued intervals are $[-K+2\epsilon-\tau, K-2\epsilon+\tau]$ and $[-K-\tau, K+\tau]$. Their widths differ by -4ϵ and their distances to M differ by 2ϵ, giving a score difference of $4\epsilon(d-1)$. Thus the block contrast is $8\epsilon(d-1)/d$ under either operator. Taking d to infinity proves that the uniform constant 8 cannot decrease for either operator. This does not identify the best constant at an individual fixed rank.

Effective geometry for signed correction. For center $z(x)$ and half-width $v(x) \geq 0$, the standardized residual is $(|Y - z(X)| - v(X))/c(X)$. Adding $hc(x)$ to every raw half-width, with constant $h \geq 0$, subtracts h from each calibration residual and its signed order statistic. The issued endpoints, including any midpoint collapse, are unchanged. For any $t \in \mathbb{R}$, choose $T \geq \max(0, -t)$; widen candidate a by $c(T+t)$ and b by cT. Their issued intervals remain unchanged. The new normalized half-width difference is $w+t$, and the largest normalized raw endpoint difference at x is $u(x)+|w(x)+t|$, where $u = |z_a - z_b|/c$ and $w = (v_a - v_b)/c$. Proposition 2 applies to these representatives. Writing $A_+ = \sup_x(u+w)$ and $A_- = \sup_x(u-w)$ yields $\sup_x\{u+|w+t|\} = \max(A_+ + t, A_- - t)$, minimized at $t = (A_- - A_+)/2$ with value $(A_+ + A_-)/2$. This proves Equation 5 when both global suprema are finite. An observed maximum is not such a certificate. Nonnegative correction does not inherit the invariance: with all outcomes zero, a raw point interval $[0,0]$ and a raw interval $[-1,1]$ both issue $[0,0]$ after signed calibration, whereas nonnegative correction leaves the latter at $[-1,1]$.

A realizable class starts from ordered reference endpoints, adds $\epsilon c(x)h(x)$ componentwise with $\|h(x)\|_\infty \leq 1$, and applies midpoint projection if needed. Projection is sup-norm nonexpansive and leaves the reference fixed, establishing the global geometry. Bounded-output neural heads are one standard implementation, not an architectural contribution here. All functions and bounds must be fixed upstream of selection.

Suppose $0 < c_{\min} \leq c(x) \leq c_{\max} < \infty$ and $\mathbb{E}(|Y| + |L_b(X)| + |U_b(X)|) < \infty$. The standardized reference residual has finite first absolute moment. The absolute value of a finite calibration order statistic is at most the sum of the absolute residuals in the pool. The issued endpoints are bounded by the raw endpoints plus $c_{\max}$ times that sum; midpoint projection does not increase this bound. Hence the reference risk is finite. Bounded candidate displacement gives finite risk for the other candidates. Unbounded response support is compatible with all these conditions.

B.2 Proof of Proposition 3

For the point candidates ϵ and zero, condition on whether the independent size-m pool contains an outcome M. Averaging its two possible calibration maxima and the independent query gives

$$D(p) = -2\epsilon + 4\epsilon(1-dp)(1-p)^{d-1}.$$

In particular, $D(0) = 2\epsilon$ and $D(1/d) = -2\epsilon$. Under $p = 0$ every observed label is zero, so a randomized rule satisfying the harm constraint updates with probability at most δ on that input. Under $p = 1/d$ the same input occurs with probability $(1-1/d)^N$. The failure probability is at least $(1-\delta)(1-1/d)^N$, and regret at least 2ϵ times this expression. Solving for a failure probability at most $\beta < 1-\delta$ proves the bound.

The input covariate is constant, so unlabelled data do not reveal p. Every label informative about p, including fit and pilot labels, counts in N. Known support and endpoints can be supplied to both rules. The lower bound concerns these two fixed actions along $\alpha = 1/d$; it is not a fixed-α, growing-m statement or a bound on arbitrary refitted models. It specializes the classical two-point testing argument.

Proof of Corollary 1. On $p \in [0, 1/d]$, the displayed risk difference $D(p)$ is strictly decreasing. At $p = 1/(8d)$, Bernoulli's inequality gives $D(p)/\epsilon \geq -2 + 4(7/8)^2 = 17/16$; at $p = 5/(8d)$, $D(p)/\epsilon \leq -2 + 4(3/8) = -1/2$. Hence roots $D(p_-) = \gamma$ and $D(p_+) = -\gamma$ exist inside this interval. Differentiation gives

$$|D'(p)| = 4\epsilon\{d(1-p)^{d-1} + (d-1)(1-dp)(1-p)^{d-2}\} \geq \tfrac{3}{2}\epsilon d$$

there, so $p_+ - p_- \leq 4\gamma/(3\epsilon d)$. Also $q(1-q) \geq 1/(16d)$ throughout the interval. The Bernoulli KL–chi-square comparison bounds either directed divergence by

$$\mathrm{kl}(p_-, p_+), \mathrm{kl}(p_+, p_-) \leq \frac{256\gamma^2}{9\epsilon^2 d} \leq \frac{32\gamma^2}{\epsilon^2 d}.$$

Let E mean that the rule chooses a. Under p_-, $\Pr(E) \le \delta$; under p_+, $\Pr(E) \ge 1-\beta$. Data processing of the laws of all N informative labels through the possibly randomized decision gives, in both directions, N times the corresponding Bernoulli divergence at least $\mathrm{kl}(\delta, 1-\beta)$ or $\mathrm{kl}(1-\beta, \delta)$. Combining the inequalities proves the corollary. Auxiliary fit labels carrying information about p are counted; the final pool remains unread.

B.3 An implementable ordinary variance-aware control

Let $h = K_a - K_b \in [-B, B]$, $v = \mathrm{Var}(h)$, and let $\widehat{D}$ be its complete U-statistic on N IID selection observations. The usual random-permutation blocking argument applies Bernstein's moment bound to $n = \lfloor N/d \rfloor$ independent block kernels, yielding

$$\Pr\{D - \widehat{D} > t\} \le \exp\{-nt^2/(2v + 4Bt/3)\}.$$

This is an established U-statistic concentration argument (Peel et al., 2010), not a new inequality technique.

For an implementable variance upper bound, partition the first $2dq$ rows into $q = \lfloor N/(2d) \rfloor$ fixed disjoint pairs of blocks. Let $W_i = (h_{2i-1} - h_{2i})^2/2$. Then $\mathbb{E}W_i = v$, $0 \le W_i \le 2B^2$ and $\mathrm{Var}(W_i) \le 2B^2 v$. Bernstein's lower tail implies, with probability at least $1-\eta$,

$$v - \overline{W} \le 2B\sqrt{vx/q} + 4B^2 x/(3q), \qquad x = \log(1/\eta).$$

Solving this quadratic in $\sqrt{v}$ gives

$$V = \min\left\{B^2, \left(B\sqrt{x/q} + \sqrt{\overline{W} + 7B^2 x/(3q)}\right)^2\right\} \ge v.$$

Use $\eta = \delta/2$ in each of the variance and mean bounds. With $c_0 = 2Bx/(3n)$, the upper confidence radius is $c_0 + \sqrt{c_0^2 + 2Vx/n}$. A union bound is valid although the variance estimate and complete mean reuse the same labels. This conservative control uses fixed paired blocks for variance; it is not the full order-$2d$ empirical variance estimator of Peel et al. (2010).

A second ordinary control uses the mean and sample variance of n independent block kernels, with radius

$$\sqrt{2\widehat{v}\log(2/\delta)/n} + 14B\log(2/\delta)/(3(n-1)),$$

by Maurer & Pontil (2009). It compares its own block mean, not the complete-U mean. Both controls use the same candidate geometry. They pay for all selection labels and block calculations; neither uses final labels or the evaluator's law parameters.

B.4 Unbounded-support comparison and accounting

The registered family uses $X \sim \mathrm{Uniform}[-1, 1]$, $c(X) = 1 + X/4$, baseline center $X/2$, and half-width $c(X)$. The alternative adds $c(X)\tanh(2X)/2$ to the center. Responses are

$$Y = X/2 + \theta c(X)\tanh(2X) + c(X)\xi/2, \quad \theta \in \{0, 1/2, 1\},$$

where ξ is standard Gaussian or a Student-t_3 variable divided by $\sqrt{3}$. Neither distribution is truncated. Both have finite first and second moments. The class has $\epsilon = 1/2$, $c_{\max} = 5/4$. The evaluator knows θ and the noise law; selection rules do not use them. There is no fitting or pilot cost for these fixed analytic candidates.

For each of the 24 distribution–calibration settings, an independent 16,384-pool reference with eight queries per pool estimates the sign of the candidate risk. Approximate Student intervals use pools as the unit and Bonferroni correction over the 24 signs. An interval containing zero leaves the sign unresolved. This reference is evaluation-only. Each of 32 selection repetitions at each budget is evaluated with 128 new calibration pools and 16 queries per pool, paired across the two candidates. The repeated selection sample, not an individual query, is the inference unit for selected-risk comparisons. Bonferroni Student intervals cover the 432 non-reference method–setting contrasts approximately; they are Monte Carlo intervals, not finite-sample coverage theorems.

The seven rules are keep, raw-score minimization, complete-U minimization, query-envelope Hoeffding, rank-envelope Hoeffding, the complete-U Bernstein control above, and the independent-block

empirical Bernstein control. Every exact tie keeps the reference. The implementation additionally requires a numerical margin 10^{-10}; this does not certify arbitrary floating-point error. All adaptive rules consult N unique labels; keep consults zero. Available labels, actual decision labels, raw diagnostic cost, complete-U cost, block variance cost and final evaluation are recorded separately. Confidence calculations are timed together and charged in full to each confidence-based rule, a conservative shared-cost accounting rather than isolated timing of each formula. Input generation and evaluation are outside decision timers. Per-query quantities that overlap are never treated as independent observations.

The complete family has 72 distribution–budget cells, 2,304 selection repetitions and 16,128 method records. A separate implementation of the decision formulas reconstructs all 16,128 objectives with maximum discrepancy 8.89×10^{-16}; independent reference-risk intervals reconstruct exactly. These are same-author computational checks, not independent replication. The complete per-cell results and all label/cost records are in the companion. No response-support bound or empirically chosen loss maximum is used.

B.5 Fitted candidates and fixed-alpha generalization

This separate synthetic development family tests whether the analytic comparison remains useful when candidate forecasts are fitted. Each replicate draws 400 fit observations from $X \sim \mathrm{Uniform}[-1, 1]$, $c(X) = 1 + 0.2X$ and

$$Y = 1.2X + \theta c(X)\tanh(3X) + 0.6c(X)\xi, \qquad \theta \in \{0, 0.35\},$$

with standard Gaussian or untruncated $t_3/\sqrt{3}$ noise. A linear least-squares fit defines the reference. The tested action adds its residual-fit coefficient for $c(X)\tanh(3X)$, clipped to $[-0.3, 0.3]$. Thus the whole input domain, not merely the observed sample, obeys the common-scale endpoint bound $\epsilon = |\widehat{\beta}|$ with $0.8 \le c(X) \le 1.2$. An ordinary unrestricted joint least-squares fit of intercept, slope and nonlinear feature provides a forecast-quality comparator for the complete bounded fitting procedure. This contrast combines its fixed baseline, sequential residual fit and coefficient clip; it does not isolate the cost of any one step. The joint fit uses the same fresh calibration but is not included among the two geometry-controlled selection actions.

Precisely, on the fitting sample let $Q = [\mathbf{1}, X]$ and $h = c(X)\tanh(3X)$. With $\widehat{b} = Q^{+}Y$, the correction before clipping is $\widehat{t}_{\mathrm{seq}} = h^{\top}(Y - Q\widehat{b})/(h^{\top}h)$; its feature is not residualized against Q. The reference and alternative centers are $[1, x]\widehat{b}$ and $[1, x]\widehat{b} + \mathrm{clip}(\widehat{t}_{\mathrm{seq}}, -.3, .3)c(x)\tanh(3x)$, both with raw half-width $c(x)$. The comparator is the joint least-squares predictor on $[Q, h]$, not a third selection action. Both selection actions use the signed residual $|Y - z(X)|/c(X) - 1$ and the finite-rank correction $q = r_{(\lceil (m+1)(1-\alpha) \rceil)}$; after correction, a crossed interval collapses to its midpoint. This specifies the fitted estimators and issuance rule without attributing their calibrated-score difference to any single fitting step. Across the 256 saved fits, the ratio $h^{\top}(I - QQ^{+})h/(h^{\top}h)$ has median .0733 (range .0609–.0900); none of the sequential coefficients reaches the clipping threshold. This is a post hoc coefficient diagnostic, not an attribution of final calibrated risk.

The sixteen cells cross two noise laws, two values of θ, $m \in \{9, 19\}$ and $N/(m+1) \in \{64, 512\}$ at fixed $\alpha = .1$. Each has sixteen independent fit, selection and final-evaluation repetitions. Every rule receives the same N selection observations, except keep, which consults no selection labels. The complete-U estimator is compared with an independent-block estimator and with 32- or 128-draw incomplete-U estimators. Random subsets may overlap as data, but are independently drawn conditional on the observed selection sample. Their upper bounds add a complete-U data-error radius at $\delta/2$ and a conditional Monte Carlo radius at $\delta/2$. Counting sampled subsets as independent data blocks would be invalid. A separate full-U Bernstein rule uses a fixed-pair block variance upper bound. All methods use identical final pools of 256 calibration sets with sixteen test queries each; neither these outcomes nor the comparison with the joint fit enters selection. Selector timers exclude the shared fit, final evaluation and process setup; a half-second incremental cap applies to each rule. No rule timed out.

Table 8: Final interval-score differences in all 16 fitted IID settings. Each setting has 16 independent fits; all five confidence rules kept the reference. The last column counts raw-rule updates and observed worse scores on finite final pools.

Noise	θ	m	N/d	Bounded Δ	Joint Δ	Raw Δ	Raw updates/ finite-final worse
Normal	0.00	9	64	-0.00127	+0.00354	-0.00108	8/1
Normal	0.00	9	512	-0.00029	+0.00131	-0.00087	9/1
Normal	0.00	19	64	+0.00036	+0.00498	-0.00045	6/0
Normal	0.00	19	512	+0.00088	+0.00307	-0.00033	7/0
Normal	0.35	9	64	-0.00266	-0.02189	-0.00453	10/1
Normal	0.35	9	512	-0.00348	-0.02123	-0.00448	10/1
Normal	0.35	19	64	-0.00318	-0.02216	-0.00357	14/4
Normal	0.35	19	512	-0.00273	-0.01497	-0.00307	13/2
Student-t_3	0.00	9	64	+0.00014	+0.00231	+0.00006	7/3
Student-t_3	0.00	9	512	+0.00054	+0.00387	-0.00014	9/3
Student-t_3	0.00	19	64	-0.00038	+0.00228	-0.00043	4/1
Student-t_3	0.00	19	512	+0.00028	+0.00237	-0.00002	7/2
Student-t_3	0.35	9	64	+0.00004	-0.01639	-0.00050	10/4
Student-t_3	0.35	9	512	-0.00069	-0.02187	-0.00095	10/4
Student-t_3	0.35	19	64	-0.00194	-0.01134	-0.00337	10/1
Student-t_3	0.35	19	512	-0.00080	-0.01618	-0.00095	5/0

In Table 8, the bounded and ordinary joint columns compare each fitted candidate with the retained linear predictor after independent final calibration. The raw column includes the empirical rule's keep decisions. The final column counts raw-rule updates and observed score worsenings on the independent finite final pools; they do not identify the population-risk sign. All five confidence-based rules make zero updates in 256 repetitions. The raw-score rule makes 139, of which 28 worsen the score measured on independent final pools. This last count concerns finite evaluation samples, not the sign of the exact population-risk difference. The bounded correction is not uniformly competitive with the unrestricted joint fit, especially when $\theta = 0.35$. Across all fixed cells the selected-minus-keep score of the raw rule averages -0.00154, a descriptive mean rather than a pooled significance test; the five controlled rules equal keep exactly. One independent program regenerates every fit, final calibration pool and final score, with maximum absolute discrepancy 8.88×10^{-16} across 256 records. Each cell's paired replicate interval and the rule-specific timing and action counts are in the companion. Zero observed updates do not establish that a calibrated action cannot improve at another sample budget or in another fitted class.

Why the fitted comparisons retain the reference. For each saved fit and rule, let r be its data radius plus any conditional Monte Carlo radius, and report $-\widehat{D}/r$. This ratio pairs quantities within the same fit; it is not a new statistic with a risk guarantee. Certification requires a value exceeding one, as well as the implementation's numerical margin. The medians and maxima across all 256 fits are $.0082, .4332$ for complete-U Hoeffding, $.0068, .7108$ for complete-U Bernstein, $.0034, .5989$ for block Hoeffding, $.0081, .2798$ for 32-draw incomplete U, and $.0021, .2191$ for 128-draw incomplete U. These are descriptive summaries, not confidence intervals or evidence that one rule has a population advantage. All rules finish within the time cap.

There are 135 negative complete-U means. At both fitted calibration sizes, Equation 12 gives $C = 5.6$. Substituting this constant for 8 in the saved Hoeffding radius still leaves every upper bound nonnegative. This is a post hoc sensitivity calculation, not a change to the registered decisions or observed label savings. Inverting the stored Bernstein radius recovers its variance upper bound; the uncapped paired squared-difference mean is recoverable in 250 fits. In the other six, the cap erases that information. Unsaved block sample variances and per-final-pool losses cannot be reconstructed from these summary records alone. We do not infer a population-risk sign from the finite-final score, nor treat the 10^{-10} decision tolerance as a roundoff certificate.

When scale functions may differ. The common-scale assumption concerns shape, not units. Multiplying a candidate's scale by a positive constant divides its standardized residuals and finite order statistic by the same constant. Query rescaling cancels this change. The issued interval is

identical, including nonnegative truncation and midpoint collapse. Thus candidates with $c_a = \lambda_a c$ can first be represented using the common shape c; the geometric bound must be checked in that representation. This is scale equivariance, not a new calibration procedure.

Different scale shapes cannot be admitted on endpoint proximity alone. For any finite rank $1 \le k \le m$, set $d = m + 1$ and $s = d - k$. Take two fixed zero-endpoint candidates and inputs u, v with $\Delta_c = c_a(v)/c_a(u) - c_b(v)/c_b(u) \neq 0$. A block with k copies of $(v, 0)$ and s copies of (u, M), $M > 0$, has

$$K_a - K_b = \frac{2Mk}{d}\Delta_c.$$

Indeed, a low query leaves $k - 1$ zero residuals, so its correction is $M/c(u)$ and its score is $2Mc(v)/c(u)$. A high query leaves k zeros, so its correction is zero and its score is $2M/\alpha$ for both candidates. Every correction is nonnegative, so either operator gives the identity. For the IID law placing probability $p \in (0, 1)$ on (u, M),

$$R_m(a) - R_m(b) = 2M(1 - p)\Pr\{\mathrm{Bin}(m, p) \ge s\}\Delta_c.$$

Both individual risks are finite for each finite M. Holding the scales fixed and letting M grow makes the contrast unbounded, despite identical raw endpoints. For example, let all scale values equal one except $c_a(v) = 1 + \eta$, with any fixed $\eta > 0$. Arbitrarily close scale shapes therefore do not supply a response-independent envelope. Additional tail or structural assumptions may permit other comparisons; this construction does not exclude them or claim impossibility for every calibration rule.

C Proofs and Local-Fit Assumptions

C.1 Proof of radius cancellation

Subtracting t from both standardized residual branches subtracts t from their maximum and every finite order statistic, regardless of score ties. The issued endpoints become $L_x - c_x t - c_x(q - t) = L_x - c_x q$ and $U_x + c_x t + c_x(q - t) = U_x + c_x q$. The same midpoint rule is consequently applied to the same pair. This proves Proposition 7. For separate tails, lower offsets add a to the lower-tail rank and upper offsets subtract b from the upper-tail rank, giving the analogous identity.

C.2 Proof of local endpoint equivalence

On the stated strict affine region, the selected residual is affine with derivative $w^\top$. The chain rule gives $J_B = A_B + g_B w^\top$. An affine map has equal outputs at two points of this region exactly when their displacement belongs to its derivative kernel. When B includes the calibration rows, the active lower-tail functional has entry $1/c_j$ on lower endpoint j, or the active upper-tail functional has entry $-1/c_j$ on upper endpoint j. Thus $h^\top g_B = -1$ and $w^\top = h^\top A_B$. It follows that $P^2 = P$ and $\ker P = \operatorname{span} g_B$. The map $v \mapsto A_B v$ from $\ker(PA_B)$ onto $\operatorname{im} A_B \cap \ker P$ has kernel $\ker A_B$. Rank-nullity gives Proposition 8. Query-only outputs need not include the active residual row, so this last dimension identity must not be applied to them without an additional argument. Neither identity is a new finite-sample estimation rate.

C.3 Local fitting and a coherent path law

We separate the executed local-fit identity from the exact-projection idealization used to establish nuisance invariance.

For one history, let $N \in \mathbb{R}^{n \times p}$ contain the declared affine nuisance trajectories and let $\Phi \in \mathbb{R}^{n \times r}$ be the learned transition basis after orthogonal projection away from their span, so $\Phi^\top N = 0$. Let $e \in \mathbb{R}^n$ be the historical transition residual, let α_0 be the amortized coefficient, and let $\widehat{\alpha}$ be the update in equation 19. The future decoder recursively maps independent innovations $Z_{1:H}$ to a path $F_x(\alpha, Z_{1:H})$. At step j, its location and positive scale may depend on x and earlier generated values, but not on future innovations.

Intuition. Orthogonalization removes the declared nuisance span, while local coefficient error transfers directly to error in the future path law.

Theorem 1 (Exact local-fit identity and path-law transfer)**.** *For every reference coefficient $\alpha_\star$,*

$$\widehat{\alpha} - \alpha_\star = (G + \lambda I)^{-1} \left\{ \Phi^\top (e - \Phi\alpha_\star)/n + \lambda(\alpha_0 - \alpha_\star) \right\}. \tag{13}$$

Holding Φ, N, and α_0 fixed, for every $\gamma \in \mathbb{R}^p$ the local fit is invariant to the declared nuisance span:

$$\widehat{\alpha}(e + N\gamma) = \widehat{\alpha}(e).$$

This invariance is stated for the exact projection $\Phi^\top N = 0$. The implementation residualizes with a ridge $\lambda_a = 10^{-3}$, fixed in every fit of Section L.2, where the coefficient ridge of equation 19 is $\lambda = 0.1$; for this λ_a $\Phi^\top N/n = \lambda_a B^\top (A + \lambda_a I)^{-1}$ with $A = N^\top N/n$ and $B = N^\top g_t/n$, so the fitted coefficients satisfy $\widehat{\alpha}(e + N\gamma) - \widehat{\alpha}(e) = \lambda_a (G + \lambda I)^{-1} B^\top (A + \lambda_a I)^{-1}\gamma$ rather than being exactly invariant. The relative deviation is governed by $\lambda_a/(\lambda_{\min}(A) + \lambda_a)$, which is small only when the nuisance Gram matrix is well conditioned against λ_a and approaches one when it is not; we do not report that conditioning. The local-fit identity holds for the executed estimator, but exact nuisance invariance does not. If $\lambda \geq 0$, $\lambda_{\min}(G) \geq \kappa \geq 0$, and $\kappa + \lambda > 0$, then

$$\|\widehat{\alpha} - \alpha_\star\|_2 \leq \frac{\|\Phi^\top (e - \Phi\alpha_\star)/n\|_2 + \lambda \|\alpha_0 - \alpha_\star\|_2}{\kappa + \lambda}. \tag{14}$$

If, in addition, $\|F_x(a, z) - F_x(a', z)\|_1 \leq L_x(z)\|a - a'\|_2$ for every z, where $\overline{L}_x = \mathbb{E} L_x(Z) < \infty$, and $\mathbb{E}\|F_x(\alpha_\star, Z)\|_1 < \infty$, then the two conditional path laws obtained with the same innovation law satisfy

$$W_1(\mathcal{L}(F_x(\widehat{\alpha}, Z)), \mathcal{L}(F_x(\alpha_\star, Z))) \leq \overline{L}_x \|\widehat{\alpha} - \alpha_\star\|_2. \tag{15}$$

In the noiseless shared-basis case $e = \Phi\alpha_\star$, full rank and $\lambda = 0$ recover $\alpha_\star$ exactly.

For the implemented decoder, let every innovation have a fixed Student-t law with $\nu > 2$, rescaled to unit variance, and let its conditional variance be $\exp(2\ell_j) + v_0$ with $v_0 > 0$. The resulting triangular recursion defines an everywhere normalized joint density on the complete future path. Its exact negative log likelihood is the sum of the conditional Student-t negative log densities and the history-difference scale terms.

Proof. Substituting $e = \Phi\alpha_\star + (e - \Phi\alpha_\star)$ into equation 19 gives equation 13. The nuisance identity follows from $\Phi^\top(e + N\gamma) = \Phi^\top e + \Phi^\top N\gamma = \Phi^\top e$. The spectral bound follows from $\|(G + \lambda I)^{-1}\|_{\mathrm{op}} \leq (\kappa + \lambda)^{-1}$. Couple the two path laws with the same Z and integrate the assumed random Lipschitz inequality; the expected coupling cost upper-bounds W_1.

For the final statement, each conditional Student-t scale is strictly positive because $v_0 > 0$. Given preceding coordinates, the map from Z_j to the next path coordinate is affine with positive slope. The complete innovation-to-path Jacobian is therefore lower triangular with positive diagonal. Sequential change of variables proves normalization and yields the stated likelihood decomposition. □

The first bound is deterministic and applies to the actual fitted basis; it does not assume that the basis is independent of the same history. A stochastic $n^{-1/2}$ identification rate requires an additional exogeneity or cross-fitting premise and is not claimed. Likewise, the theorem establishes a coherent model and a conditional coefficient-error bound, not identification of a population repair direction or empirical superiority. The reference $\alpha_\star$ is arbitrary unless an additional data-generating assumption specifies it. Nuisance invariance is not invariance to recomputing the learned basis from a changed history. An innovation-uniform Lipschitz constant is not automatic for a heteroscedastic Student-t decoder; the integrable random constant is an additional assumption, not an established property of the fitted network. The Wasserstein comparison is between two model laws, not between either model and the true future law.

For $\lambda > 0$, the fixed-basis residual-to-coefficient sensitivity has the standard ridge bound

$$\|\widehat{\alpha}(e + \delta) - \widehat{\alpha}(e)\|_2 \leq \frac{1}{2\sqrt{\lambda}} \frac{\|\delta\|_2}{\sqrt{n}}.$$

Indeed, with $X = \Phi/\sqrt{n}$, the singular values of $(X^\top X + \lambda I)^{-1} X^\top$ are $s/(s^2 + \lambda) \leq 1/(2\sqrt{\lambda})$. At the implemented $\lambda = 0.1$, unit root-mean-square residual perturbations therefore have coefficient sensitivity at most 1.5811. Ill conditioning does not make this fixed-ridge operator unbounded. It can nevertheless defeat the intended nuisance suppression; changes that also recompute the basis fall outside this bound.

C.4 SELECTED-ACTION RECALIBRATION

One sufficient construction calibrates the final interval *after* its action and score map are selected. Under the conditional exchangeability assumption below, this includes center shifts, locally scaled intervals, and parent-pool fallbacks without requiring containment of a common baseline. This is not a necessity result: selection-adjusted constructions can use other orderings. Chronology alone does not establish exchangeability.

Pipeline and filtration. One run fixes the representation, cells, groups, candidate templates, target level, and support rules before outcomes assigned to later roles are read. The six chronological roles define the filtration

$$\underbrace{\mathcal{F}_{\mathrm{con}}}_{\substack{\text{construction}}} \subseteq \underbrace{\mathcal{F}_{\mathrm{geo}}}_{\substack{+\text{ geometry fit}\\ \text{scales and score maps}}} \subseteq \underbrace{\mathcal{F}_{\mathrm{adm}}}_{\substack{+\text{ acceptance}\\ \text{outcomes}}} \subseteq \underbrace{\mathcal{F}_{\mathrm{sel}}}_{\substack{+\text{ selection}\\ \text{outcomes; policy freeze}}} \subseteq \underbrace{\mathcal{F}_{\mathrm{cal}}}_{\substack{+\text{ final calibration}\\ \text{outcomes; ranks only}}} \longrightarrow \begin{array}{c}\text{deployment}\\ \text{scores}\end{array} \tag{16}$$

All candidate geometry, including a studentizing scale and the identity of a parent pool, is $\mathcal{F}_{\mathrm{geo}}$-measurable. Acceptance and selection may use their own outcomes but cannot read final-calibration outcomes. At the end of selection, the policy $\widehat{\pi}$, its row-to-action map, every emitted score function $s_{\widehat{\pi},z}(x,y)$, and every calibration-pool identity are $\mathcal{F}_{\mathrm{sel}}$-measurable. Final calibration estimates only the rank of the score of the *selected* emitter. No endpoint is clipped, rescaled, or replaced after that rank is observed.

Assumption. Fix a cell z that the frozen policy emits. Let $S_i = s_{\widehat{\pi},z}(X_i, Y_i)$ for its n_z final-calibration examples and $T = s_{\widehat{\pi},z}(X, Y)$ for a deployment example. We require

(E) Conditional on $\mathcal{F}_{\mathrm{sel}}$, the vector $(S_1, \ldots, S_{n_z}, T)$ is exchangeable.

The premise is local to the selected cell and score map. It is not implied by chronology, and a dependence-aware replacement theorem is required when serial dependence invalidates it.

Intuition. Selecting a score map first and recalibrating that exact map afterward recovers ordinary split-conformal validity for the issued interval.

Theorem 2 (Selected-action conformal validity)**.** *Let $k_z = \lceil (n_z+1)(1-\alpha) \rceil$. If $k_z \le n_z$, let $q_z = S_{(k_z)}$ and emit $C_z(x) = \{y : s_{\widehat{\pi},z}(x,y) \le q_z\}$. Under (E),*

$$\mathbb{P}\{Y \in C_z(X) \mid \mathcal{F}_{\mathrm{sel}}\} \ge 1-\alpha.$$

The result holds for a selected center shift, a geometry-fit studentized score, or a selected parent-pool emitter without a multiplicity correction over the candidate family. It holds because selection precedes the independent final calibration, not because the selected interval contains another interval. If $k_z > n_z$, the protocol emits no $1-\alpha$ claim and follows its support rule, which is frozen before final-calibration outcomes are read.

Proof. Conditional on $\mathcal{F}_{\mathrm{sel}}$, the selected policy, selected score map, pool identity, and rank are fixed. Assumption (E) makes the rank of T among $S_1, \ldots, S_{n_z}, T$ uniform after random tie breaking. Hence $\mathbb{P}\{T > S_{(k_z)} \mid \mathcal{F}_{\mathrm{sel}}\} \le 1 - k_z/(n_z+1) \le \alpha$. The argument is applied only to the selected score map, so no union over candidates appears. Geometry-fit scaling and parent-pool selection are already fixed under the conditioning sigma-field. □

D FULL EXPERIMENTAL RESULTS

D.1 POINT-ACTION ACCEPTANCE

The component study uses 17,379 hourly Bike Sharing observations. The first 50% construct a baseline and an event-state candidate, the next 20% accept and select an action, the next 15% calibrate intervals, and the final 15% are held out from fitting. Five training-only displacement caps and two strengths give ten candidates. Seven absolute-loss tests cover the aggregate, body, high-loss tail, and four chronological quarters, and the confidence allocation covers all 70 candidate-group tests.

This component uses an IID empirical-Bernstein acceptance bound; it is not the blocked dependence theorem. The accepted candidate uses the training-median displacement cap, 17.489, at strength 0.5; its largest acceptance upper bound is -0.111. On the held-out evaluation role, tail MAE falls from 388.582 to 380.015. More aggressive edits obtain larger average design gains but are rejected. Table 9 reports that evaluation.

Table 9: Held-out Bike evaluation after simultaneous point-action acceptance. Tail rows are the baseline's training-defined highest-loss decile. The acceptance rule protects absolute loss; asymmetric losses are report-only.

Metric	Fixed baseline	Accepted action
MAE	98.135	94.867
RMSE	154.302	150.464
Tail MAE	388.582	380.015
Underprediction loss, weight 2	187.538	182.247
Underprediction loss, weight 5	455.746	444.386

The selected candidate passes the 70-test absolute-loss contract. It does not pass the broader 210-test contract that also protects the two asymmetric losses; the largest upper bound changes from -0.111 to 4.069.

Table 10 is the direct comparison in the first development window: every method scores the same rows against the same fixed point centers at the same target, so the columns are read against one another rather than against published numbers.

Table 10: Identical-sample comparison in the development window. Every fixed-center interval repair number here is the earlier fixed-center candidate, not a routed system. Coverage target is 0.90. The post-calibration width-projection ablation uses a final-calibration-dependent projection that is not covered by our restricted theorem.

Method	Coverage	Score	Width	Weak $<$ .88	Worst
calibrated baseline	.8897	209.55	141.48	81	.3513
split CP	.8713	261.60	121.88	97	.4086
normalized split CP	.8750	210.33	139.45	92	.3513
CQR-HGB	.8776	260.69	194.43	115	.7049
ACI	.8997	1009.34	959.54	62	.7033
EnbPI adapter	.7989	179.72	91.23	187	.4139
SPCI adapter	.7189	200.01	76.78	220	.4438
CPTC adapter	.8715	178.26	117.34	122	.7258
conformal PID adapter	.9214	181.43	113.38	35	.7767
post-calibration width projection	.9010	161.30	102.79	54	.7817
fixed-center interval repair candidate	.9025	161.70	103.73	49	.7893

Mean width hides how heavy the rowwise tails are, so Table 11 reports the width ratio to the calibrated baseline at the median and at three upper percentiles in both windows.

Table 11: Evaluation-row width ratios of the fixed-center candidate relative to the calibrated baseline. These tails are not visible from mean width alone.

Window	Median	P90	P95	P99	Maximum	Rows $>$ 10
First	0.760	1.412	1.953	8.049	44.11	4,919
Second	0.850	1.709	2.233	12.855	110,187.78	9,204

Table 12 repeats the same identical-row comparison in the second inspected window, on a disjoint time range and the same eleven methods, which is how we check that the ordering is not an artifact of one window.

D.2 Comparator fidelity

Not every comparator can be run at the same fidelity, so Table 13 states, for each row of the two comparison tables, how faithful the implementation is and what boundary that fidelity imposes on reading it.

Table 12: Identical-sample comparison in the second inspected development window, again scoring the earlier fixed-center candidate and not the routed instance. Coverage target is 0.90.

Method	Coverage	Score	Width	Weak $< .88$	Worst
calibrated baseline	.9236	227.91	130.29	52	.3190
split CP	.9105	305.79	113.26	56	.3522
normalized split CP	.9132	227.36	124.40	71	.3049
CQR-HGB	.9085	288.14	198.03	73	.6369
ACI	.9205	214.11	145.14	36	.6623
EnbPI adapter	.8289	210.02	98.93	171	.3468
SPCI adapter	.7474	242.88	82.61	223	.4681
CPTC adapter	.8819	211.90	137.51	105	.6018
conformal PID adapter	.9196	209.46	122.83	45	.7433
post-calibration width projection	.9393	168.49	126.35	24	.7780
fixed-center interval repair candidate	.9541	166.19	123.64	2	.8438

Table 13: Fidelity labels for every row of Tables 10 and 12. "Official adapter" means that the released interval class ran on fixed centers and identical rows; it is not native end-to-end parity.

Method family	Fidelity	Direct-table boundary
calibrated baseline	exact algorithm	reference row for every width ratio
split / normalized split	exact algorithm	same centers and rows
CQR-HGB	canonical reproduction	declared quantile regressor
ACI	canonical reproduction	fixed gain, origin-safe feedback
EnbPI / SPCI	official adapter	no native point ensemble
CPTC	official adapter	fit-only state proxy
conformal PID	official adapter	not error-quantified parity
post-calibration width projection	development ablation	theorem-incompatible projection
fixed-center interval repair candidate	frozen candidate	inspected development panels

Native end-to-end evaluations, point-revision methods, relational conformal prediction, and multivariate prediction sets belong in separate tables because they change the center, input information, training budget, or output geometry. The anonymous fidelity index provides a one-to-one mapping from every descriptive table label to the immutable result key used by the validator.

Error-quantified conformal inference under a fixed center. Error-quantified conformal inference updates the level by an error-quantified gradient step (Wu et al., 2025). It has no row in Tables 10 and 12, and earlier adapters of ours carrying that label are not evidence about it: they execute ACI, a quantile tracker, or a simplified proportional-integral controller. The implementation used here transcribes the paper's defining recursions and separately imports the interval classes of the official repository, and a synthetic self-test passes twelve machine-verified checks, including step-exact recursion identities. The self-test also records three places where the released code differs from the equations as printed: the official code applies the error-quantification term transiently, whereas the defining update as written accumulates it; the official integral variant integrates only the error-quantification term while keeping the binary term single-step; and the official cutoff variant fixes its trailing window at 100 and ignores its burn-in argument. We therefore make the official-repository tier primary and retain the literal-equation tier only as a report under this caveat.

The prospectively specified real-data adapter was then executed with fixed centers on each 677,376-row development panel. Each panel contains 27 dataset-forecaster cells. The learning rate was selected inside each cell from $\{1, .5, .1, .05\}$ using only its selection role. In the first panel, 10 of 27 cells pass the selection coverage gate, 17 fail it, and 9 cells produce an invalid evaluation interval; only 497,664 rows are emitted, with 8 reversed evaluation intervals and at least one non-finite endpoint. In the second panel, the corresponding counts are 5 passes, 22 failures, 8 invalid cells, 433,152 emitted rows, and 6 reversed intervals. Thus full-surface width and interval score are undefined, and this method is excluded from the numeric rankings. We neither sort endpoints nor substitute a learning rate after evaluation. This is a method-failure certificate for this fixed-center interface, not an end-to-end comparison with the forecasting pipeline of Wu et al. (2025), whose own reported results we neither reproduce nor dispute.

Table 14: Frozen final-output crossing census before uniform rearrangement. Rates are paired domain-stratified cluster means; intervals are 10,000-draw cluster-bootstrap intervals for foundation minus classic.

Panel	Foundation	Classic	Difference	95% interval
A	.5124%	.0936%	.4188 pp	[.2258,.6360] pp
B	.3003%	.0453%	.2550 pp	[.1301,.3889] pp

Official point-revision evidence. PIR (Liu et al., 2025) is kept outside the interval table. At horizon 96, its official runs reduce PatchTST (Nie et al., 2023) task-level MAE from 0.415774 to 0.400040 on ETTh1 and from 0.219079 to 0.208723 on Weather. Under a chronological design, acceptance, and evaluation split with a full 96-origin embargo, ETTh1 MAE changes from 0.450221 to 0.442703, but no candidate passes the origin-block sensitivity, so the dependence-aware decision keeps the baseline. On Weather, full PIR changes MAE from 0.203294 to 0.191426, but protected channels regress and every candidate fails even the IID sensitivity. These are informative point-action comparisons, not conformal interval results.

D.3 Uniform rearrangement of the frozen calibrated field

We tested whether replacing the local closed-form coefficient with learned amortization over a pretrained representation removes the remaining model gap. Before training, we froze Chronos-T5-small, 64 joint sample paths, a 134,922-parameter attention-conditioned orthogonal-innovation head, a Student-t decoder, three seeds, three folds, nine domains, and a later split-CQR calibration role. The candidate is compared with the local-ridge history-local fit and the native Chronos path law on identical rows. Dataset-series is the analysis unit; neither calibration rows nor evaluation outcomes enter model selection. The joint-path energy comparison is well defined, but a later preregistered width decomposition added an ordered-endpoint guard that the historical calibration code lacked.

Before any rescoring, we froze one deterministic rule for all 34 final calibrated surfaces:

$$L^*(x) = \min\{L(x), U(x)\}, \qquad U^*(x) = \max\{L(x), U(x)\}.$$

This is the two-quantile case of monotone rearrangement (Chernozhukov et al., 2010). It reads neither outcomes nor method identity and changes no row, radius, model, seed, or gate. It is a borrowed geometry repair, not a new conformal procedure or monotone-head theorem. The native endpoint pairs are ordered. Crossings arise only after signed CQR shrinkage: a negative calibration radius whose magnitude exceeds the native half-width reverses the final pair. Table 14 gives the crossing census before rearrangement.

Overall, 31 of 34 surfaces, or 91.2 percent, contain a crossing, with 2,781 crossed rows and an equal-surface row rate of .2295 percent. The family contrast uses the ten fixed foundation systems and two classic forecasters; the two internal models and hybrid are excluded. Both confidence intervals are above zero, establishing the predeclared contrast in crossing prevalence. This is fixed-method-population inference on opened-development panels, not row-IID inference, model-population inference, or evidence that the native foundation outputs cross. For any two-endpoint interval pipeline, rowwise minimum and maximum provide a deterministic, outcome-blind ordering step with constant work per row; we recommend applying it before interval evaluation.

The rearranged width decomposition rejects the proposed width explanation: width share, score-width tradeoff, coverage-standardization stability, and leave-one-domain-out stability pass zero of four gates on each panel. After coverage standardization, the fixed-method-population foundation-minus-classic score advantage remains positive on both panels; its fractional narrowing is -0.0164 on panel A and $+0.0496$ on panel B. These development data do not support the conjecture that the score advantage is purchased by wider intervals, and no universal family-width claim is retained. Table 15 gives the hybrid against local ridge after rearrangement.

Rearrangement restores an interval-score gain: the two-panel median improvement over local ridge is 6.268 percent. Promotion remains rejected. Only interval noninferiority, strict interval improvement, coverage, and integrity pass; energy noninferiority, native-Chronos noninferiority, top-four rank, and the 1.05 width-ratio criterion fail. Only four of eight criteria pass, so the model is not promoted. An independent implementation reconstructs 31,503 crossing quantities, 59,906 width quantities, 1,890

Table 15: Hybrid versus local ridge after uniform rearrangement. Positive numbers favor the hybrid. Intervals use the dataset-series cluster.

Panel	Interval-score gain	Energy gain	Width ratio	Field rank
A	7.08% [4.26,9.61]	-3.72% $[-5.35, -2.13]$	1.641	11/15
B	5.46% [2.93,7.77]	-4.58% $[-6.42, -2.85]$	1.590	11/15

Table 16: Frozen Chronos-2 three-arm comparison at 80% nominal coverage. IS is calibrated interval score; ES is energy score. Lower IS, width, and ES are better.

Panel	Method	Coverage	Width	IS	ES
A	Native Chronos-2	.7918	2.9652	5.1706	1.2341
A	Matched quantile head	.8046	4.3109	6.4326	1.3287
A	orthogonal-innovation path head	.8078	5.2986	7.3870	1.4639
B	Native Chronos-2	.8008	2.8358	5.0440	1.2046
B	Matched quantile head	.8165	4.0861	6.0044	1.2675
B	orthogonal-innovation path head	.8131	4.9753	7.1221	1.4113

leave-one-domain-out transfer quantities, and 238 head-comparison quantities, 93,537 in total, with maximum numerical discrepancy zero.

D.4 Matched-backbone Chronos-2 head transfer

We next test the representation-confounding explanation directly. Before any fit or score was opened, we froze the Apache-2.0 Chronos-2 checkpoint, its 768-dimensional last output-patch state, nine datasets, three folds, three seeds, and three arms: native Chronos-2, a 163,358-parameter direct quantile head, and the 163,594-parameter orthogonal-innovation path head. That head is the orthogonal-innovation path head of Appendix D.3, first fitted on frozen Chronos-T5-small states and here reading Chronos-2 states; it is not the standalone model of Table 38, which trains its own history encoder. The relative trainable-parameter gap between learned heads is 0.144 percent. Both marginal-quantile emitters use the same training-role native-PIT Schaake template for path scoring; history-local fit retains its intrinsic Student-t path law. Split conformal calibration and the same outcome-blind endpoint rearrangement are applied after fitting.

Relative to native Chronos-2, the path head worsens IS by 42.87 and 41.20 percent and ES by 18.62 and 17.16 percent on panels A and B. Relative to the matched quantile head, it worsens IS by 14.84 and 18.61 percent and ES by 10.17 and 11.35 percent. Every paired domain-stratified dataset-series bootstrap interval excludes zero in the unfavorable direction. Thus no metric family passes and the head-transfer study fails all eight predeclared criteria. The conclusion is narrow: that head's calibrated interval-score gain on Chronos-T5-small does not transfer to Chronos-2 under this frozen interface. It does not establish that lightweight adaptation of Chronos-2 is impossible. An independent implementation reconstructs 55,136 quantities over 55,080 scored rows with maximum numerical discrepancy zero.

A preregistered follow-up asks whether the null survives a head initialized at the native forecast. The gated residual head returns the native quantiles plus a learned residual scaled by a gate and by an amplitude initialized at zero, so at initialization it is exactly native Chronos-2; it has 163,472 trainable parameters against the 163,358 of the matched quantile head. On the nine development domains it passes all eight preregistered gates: relative to native Chronos-2, calibrated interval score improves by 5.73 and 6.10 percent (paired cluster intervals [4.16, 7.30] and [4.25, 8.01]) and energy score by 10.37 and 11.27 percent ([8.55, 12.27] and [9.38, 13.31]) on panels A and B, with calibrated coverage 0.7913 and 0.8102 against native 0.7918 and 0.8008; relative to the matched quantile head the gains are 24.22 and 21.12 percent in interval score and 16.76 and 15.67 percent in energy. An independent reconstruction of all 55,080 scored rows reproduces every quantity with zero discrepancy. Two confirmation surfaces were then opened once each, under the frozen heads and separate preregistrations, and both are reported here. Two public collections held back from this head's development, seven weekly influenza-like-illness series and five half-hourly Elec2 electricity-market series (twelve clusters), had been used in earlier project work and are therefore held back rather than untouched; they return one of four comparisons in favour: energy on panel A improves by 1.66 percent ([1.09, 2.31]), while interval score on panel A (+0.39 percent, [-0.87, 2.02]) and both

metrics on panel B (-3.60 and -3.12 percent, [-7.42, 2.77] and [-6.07, 2.24]) have intervals containing zero. The public surface is decisive against the head: on 293 daily series of the ICML 2026 TIME benchmark (Qiao et al., 2026), opened after the mixed result and preregistered to be reported whatever it said, native Chronos-2 is better on both metrics and both panels with every interval excluding zero (interval score -6.999 percent [-7.553, -6.457] and -6.318 percent [-6.880, -5.744]; energy -2.248 percent [-2.528, -1.977] and -1.815 percent [-2.120, -1.510]), at equal calibrated coverage and larger width. A second preregistered variant projects raw residual intervals to native width during training. It again passes all eight development gates (calibrated interval score better than native by 5.261 and 5.022 percent, energy score by 11.26 and 11.12 percent) and again loses on the same 293 public series (interval score -3.829 percent [-4.204, -3.456] and -3.399 percent [-3.783, -3.019]; energy -0.7383 percent [-0.9644, -0.5143] and -0.3725 percent [-0.6245, -0.1134], all four intervals excluding zero against the head). This is a sensitivity result on an already opened surface, not a second independent confirmation or an identified causal contribution of width: projection also changes optimization, and final calibration can widen the projected intervals. These are archived numerical outcomes, subject to the calibration and sampling qualifications below. The conclusion is narrow: a native-preserving residual head improves on native Chronos-2 inside the development panels and does not transfer, so the earlier transfer failure stands and no cross-domain adoption claim is licensed.

Validity boundary of the head-transfer experiments. The confirmation pipeline has an outcome-availability defect for the influenza series and the TIME benchmark's ECDC COVID-19 series: with a sixteen-step horizon and eight-step stride, the final eight targets of the last calibration origin occur after the first evaluation cutoff on each panel. In addition, calibration pools outputs from three folds without a fold key, repeating target observations instead of independently calibrating each fitted emitter. The printed interval metrics remain numerical descriptions of that pipeline, not evidence of causal deployment coverage or valid per-emitter split calibration. Raw energy scores are computed before calibration and do not use those radii, so this particular defect does not alter them. The corrective replay below changes calibration only, without claiming verified deployment-time label release.

The comparison is interface-controlled: histories and targets are resampled to a common grid and clipped in standardized units. It is not an evaluation of each dataset's unmodified native task. The native sampling adapter uses empirical probability-integral-transform values without rank uniformization; its samples need not preserve the advertised quantile marginals. Energy results therefore compare the implemented sampled laws, not a verified marginal-preserving copula. The residual head also receives native quantiles, starts at the native forecast, and adds an energy term to its training loss; the direct head does not share all three choices. Near-equal parameter counts do not isolate an architectural contribution. These distinctions limit both positive development and negative transfer interpretations.

Chronological recalibration of the saved forecasts. We subsequently fixed a corrective protocol on the three already opened forecast archives. Each emitter fitted with a fixed seed and fold is calibrated separately; only labels whose interpolation support ends by the first evaluation origin are retained. The finite-sample quantile rank is not capped at the number of available labels; insufficient support would produce an unbounded interval. We report both signed-radius and nonnegative-radius rules, alongside the historical pooled rule, emitter separation alone, and additional calibration delays of one and two original horizons. All 36 combinations of archive, panel and protocol complete without an empty or infinite-radius pool. The isolated original calibration functions reproduce six previously reported panels with zero numerical discrepancy; historical results are retained unchanged.

Table 17 summarizes the zero-delay nonnegative rule, with the full six-rule comparison archived. The small reserve comparison remains unresolved, while both learned heads remain worse than the native baseline on the supply-chain and health collection. All six rules retain this latter negative direction on both panels; the interval-score gap cannot be attributed solely to fold pooling or the corrected label-timing defect. The width-projected head can reduce the descriptive weak-series count while still worsening score and width, so this does not establish joint superiority.

This analysis assumes immediate availability at the source observation index, not verified publication timestamps. Extra delays affect calibration labels only; saved model histories, resampling and marginal laws remain unchanged. The paired bootstrap uses 1,000 draws and holds fitted models and calibration radii fixed. An item-block sensitivity groups the 293 public series into 109 entities

Table 17: Corrective recalibration with available labels and a nonnegative radius. Positive interval-score gain favors the learned head. H/N denotes head/native. Weak counts use per-series coverage below 0.78, not the earlier weak-group definition. Paired 95% intervals resample series within datasets.

Saved head and surface	Panel	IS gain (%) [95% CI]	Coverage H/N	Weak H/N
Gated, 12 reserve series	A	$+0.36$ [-0.92, 1.94]	.8207/.8258	3/3
	B	-3.59 [-7.48, 2.71]	.8170/.8320	4/2
Gated, 293 public series	A	-6.49 [-6.99, -6.00]	.8197/.8200	48/46
	B	-5.87 [-6.35, -5.37]	.7911/.7900	108/109
Width-projected, same 293	A	-3.52 [-3.85, -3.18]	.8190/.8200	46/46
	B	-3.14 [-3.47, -2.80]	.7912/.7900	106/109

Table 18: Nine-domain descriptive data census used to interpret the transfer null. Tail is the Hill exponent of positive absolute increments; burst is the conditional-to-marginal high-increment risk ratio. Boundary is an observable pile-up proxy, not a censoring estimate.

Data	Tail	Zero	Burst	Boundary	Common	Lifecycle	Unique	Scale CV
ECL	1.177	.0168	10.05	.0170	.516	.575	.134	1.488
ETTh1	1.529	.0055	9.70	.0005	.342	.525	.060	.784
ETTh2	3.117	.1618	5.08	.0005	.463	.892	.106	.556
ETTm1	1.773	.0031	7.07	.0005	.431	.496	.042	.933
ETTm2	2.802	.0236	3.95	.0241	.562	1.066	.084	.625
Exchange	2.495	.0000	2.76	.0018	.490	1.116	.818	.697
Solar	2.168	.6269	9.09	.6271	.936	.103	.107	.268
Traffic	1.482	.0123	6.50	.0126	.583	.142	.221	.407
Weather	0.701	.1904	13.87	.2015	.448	.479	.221	1.594

and retains the negative intervals, but neither resampling scheme accounts for common calendar shocks or fitting and calibration uncertainty. These are corrective comparisons on opened data, not new confirmation, native-task validation, or an unconditional coverage guarantee. Table 18 gives the descriptive nine-domain census, with tail exponent and burst rate, that we use to interpret the transfer null.

Traffic is the only dataset on which history-local fit improves calibrated interval score over native Chronos-2 on both panels. Energy improves on six of nine datasets per panel, but the pooled result is reversed by large losses, especially on zero-inflated Solar. Across only nine domains, the largest descriptive Spearman relation is between unique-value fraction and panel-A energy gain ($\rho = .70$); zero fraction and panel-A interval gain have $\rho = -.55$. These exploratory associations are neither multiple-testing adjusted nor causal. They motivate residual-to-native shrinkage and richer relational states, not a retained mechanism claim.

E Reproducibility and Evaluation Scope

E.1 Information and evaluation roles

Fitting, internal calibration, action selection and final calibration use separate roles in the controlled comparisons. Evaluation labels report outcomes and never choose a deployed candidate. IID examples use independent draws; chronological examples instead use their stated history and release assumptions. Sample identity includes dataset, series, issue, horizon and timestamp. Comparator capacity and feedback rights are reported rather than inferred from a shared input name.

E.2 Predeclared development criteria

The following checks were fixed before evaluation in each time window, and they govern the earlier fixed-center candidate in Appendix D, not every later model comparison: coverage at least 0.90; interval score no worse than baseline; mean width ratio at most 1.02; weak-group count reduced by at least 10%; worst-group coverage improved by at least 0.03; forecast-shift score and weak groups

Table 19: Bounded identical-row comparisons on 153 series. Positive median interval-score differences favor the named baseline; intervals use series clusters.

Baseline	Local minus baseline	95% interval
DeepAR Student-t	$-.037$	$[-.607, .166]$
N-HiTS quantiles	$+1.822$	$[1.431, 2.468]$

no worse; repair rate at least 0.01; fallback rate at most 0.50. The development artifacts label fallback as abstention and still emit the baseline interval; this is not non-emission. The weak-group reporting floor is 0.88 and minimum support is 100.

E.3 HISTORY-LOCAL FIT MATCHED-MODEL CONTRACT

The fitting-time experiment uses three fixed seeds and three domain folds over the nine public datasets. In each fold, three domains are excluded from training and scored once. The history-local fit and the amortized control have 52,986 trainable parameters and share the complete transition network, residualized basis, Student-t decoder with five degrees of freedom, variance floor 0.01, initialization, optimizer, epochs, histories, targets, and Monte Carlo scoring draws. The sole intervention is replacing the encoder-produced coefficient by equation 19. Gaussian and chronological autoregressive comparators retain their own likelihoods.

Each execution stores the trained checkpoint, row-level metric archive, source hashes, parameter counts, device, environment, and domain identities. The aggregate binds all 36 task reports and their prediction hashes. The primary criteria are stated in Appendix L.2.

The reported model is a decoder retry, and its history matters for how the result should be weighed. A first version with a Gaussian decoder, otherwise identical, lowered energy against the amortized control in every seed but was rejected under its own frozen contract, which also made three further checks fatal: against the chronological autoregressive model one seed's clustered interval reached $+0.00877$; its Weather likelihood was unstable, with negative log-likelihood 52,860 to 105,475 against 86 to 137 for the autoregressive model; and one seed's width ratio, 1.0277, exceeded a 1.02 safeguard. A ruling made after that rejection treated those three checks as reported diagnostics rather than gates, kept the four primary criteria, and allowed at most two retries that could change only the decoder law and add a variance floor, with the encoder, basis, local update, splits, seeds, controls, optimizer, budgets and evaluation rows frozen. The Student-t retry reported here passed every primary criterion and every diagnostic (Weather likelihood within ten times the autoregressive model's, width ratio at most 1.05, coverage change at most 0.035), so the second retry was not run. All nine domains had already informed development, so the result is a matched development test rather than an untouched confirmation.

The ten-seed support study adds seven seeds without changing the three-seed decision; its aggregate binds all 42 added fold-variant executions and their prediction hashes. Table 20 reports every seed. For each metric, we resample ten seed-level relative gains with replacement 20,000 times, compute each resample's median, and report its 2.5 and 97.5 percentiles. The distribution is discrete, so a percentile can equal an observed seed gain, as does the energy upper endpoint. The interval-score median is the average of 6.716990 and 6.924862 percent; five additional seeds exceed this median. This is conditional on the same datasets, not uncertainty over new domains.

The component study binds 27 trained ablation executions and reuses, rather than retrains, the no-local-coefficient control. It identifies the transition basis and history-local coefficient as the largest contributors: removing them reduces interval-score advantage by 8.52 and 4.76 percent. A Gaussian decoder costs 1.84 and 3.99 percent. Removing the floor costs 0.05 and 0.67 percent without a stable three-seed effect, so it is a numerical safeguard rather than a load-bearing component. The external comparison uses 1,836 identical row keys and 153 dataset-series clusters. An independent replay reproduces every target, context transform, and score, and records the Weather tail concentration that makes the untrimmed mean nonrepresentative. A separate analysis-unit reconciliation shows why the cluster definition matters. Treating all 1,836 rows as independent gives a DeepAR difference of -0.150 with interval $[-0.251, -0.011]$, but every dataset-series contributes 12 chronological origins. Replaying the predeclared dataset-series bootstrap gives -0.037 with interval $[-0.607, 0.166]$,

Table 20: All ten matched-development seeds. Gains are percent reductions in raw scores, not calibrated gains. The first three seeds form the primary experiment.

Seed	Energy gain (%)	Interval-score gain (%)	Energy domain wins
20260827	2.2326	4.5277	6/9
20260828	4.9705	4.7617	6/9
20260829	3.4727	6.7170	6/9
20260830	4.6716	7.0956	8/9
20260831	2.7595	4.7467	9/9
20260832	4.9984	6.0291	6/9
20260833	4.3136	9.6251	6/9
20260834	6.8521	10.3188	5/9
20260835	4.8177	9.2770	6/9
20260836	3.5623	6.9249	5/9

exactly reproducing Table 19. The row-independent interval is therefore diagnostic only. All three additions use opened development data only.

E.4 Held-out evaluation

The complete candidate, parameter-matched control, six checkpoints, public source records, deterministic TSF adapter, chronological calibration and evaluation roles, per-horizon split-conformal rule, series-level analysis unit, bootstrap, and nine conjunctive criteria were frozen before source acquisition. The experiment opened all three collections once and retained every one of the 538 eligible series. Table 21 averages the three frozen seeds within each series. Positive loss reductions favor history-local fit.

Table 21: Complete held-out result. IS is interval score, WR is the candidate-to-control width ratio, Cov reports candidate/control coverage, and the last two columns report candidate/control subgroup values.

Domain	Series	Energy ↓	IS ↓	WR	Cov	Weak	Worst deficit
FRED-MD	107	$+2.70\%$	$+17.11\%$	1.101	.765/.771	15/13	.099/.096
Pedestrian	66	$+4.58\%$	$+6.11\%$	.952	.822/.822	8/8	.164/.159
Tourism Monthly	365	-0.15%	-1.37%	.942	.816/.807	5/5	.175/.204

The equal-domain series bootstrap gives a pooled candidate-minus-control energy difference of $-.01944$ with 95 percent interval $[-.03012, -.00973]$ over 100,000 draws. Four of nine criteria pass: the pooled energy interval, median domain interval-score improvement, maximum absolute coverage change, and the per-domain candidate coverage floor. The every-domain energy, every-domain interval score, width-ratio, weak-group, and worst-deficit criteria fail. An independent program exactly reproduces all reported values and verifies one opening, three domains, six model tasks, and the prohibition on a second opening. No failed domain, series, seed, horizon, or subgroup is removed.

E.5 Frozen N-HiTS encoder-transplant test

The transfer test uses the frozen official-class N-HiTS lower, median, and upper paths as the forecast representation. Its candidate replaces an amortized mechanism coefficient by a closed-form fit over the latest eight strictly earlier training origins. Candidate and control have 83,562 trainable parameters, identical initialization, and the same transition law, decoder, optimizer, folds, seeds, and 1,836 selection rows. Selection outcomes never enter the fit. The first attempt pools the coefficient over all horizons. The second changes only the residual-scale anchor. The final attempt returns to the first scale and changes only the coefficient estimator, solving it by horizon before applying a fixed $[0.25, 0.50, 0.25]$ smoothing kernel. Table 22 gives all three attempts against their amortized controls.

All 54 trained tasks pass the sample-identity, disjoint-fold, parameter-identity, initialization, finite-output, and information-flow checks. The three null results are retained under their original decision

Table 22: All frozen N-HiTS transfer attempts. Positive energy and interval-score changes favor the transplant over its amortized control. The clustered-energy column counts seed intervals entirely below zero. A positive final column favors plain N-HiTS. No attempt passes the four-part gate.

Attempt	Energy improvement by seed (%)	Clustered energy	Median IS gain (%)	Domain wins	Difference from plain N-HiTS [95% interval]
pooled coefficient	$+.013, +.090, -.460$	0/3	$-.299$	5/6/3	$+.207$ $[+.151, +.261]$
scale-anchored	$-1.804, +.274, -.496$	0/3	-1.583	4/5/6	$+.215$ $[+.170, +.244]$
horizon-local	$+.249, +.007, -.665$	0/3	$-.482$	5/5/5	$+.223$ $[+.175, +.280]$

rule. They do not establish the successful fitting-time intervention in Table 38 as a backbone-agnostic adapter, nor do they isolate which representation or decoder property caused the failure.

E.6 Frozen shared probability-law stress test

A separately frozen extension tests whether fitting-time identification survives joint probabilistic training on a proprietary daily record of fundraising campaigns provided by an industry partner for research use: funding raised in US dollars and investor count, each at horizons of 1, 7, 14 and 30 days, give eight target-horizon domains. The candidate uses a shared causal dilated convolutional trunk, nuisance-residualized transition basis, history-local ridge coefficient, and a zero-change hurdle with signed softplus-Student-t magnitudes. Its shared amortized control and eight isolated controls have the same 32,655 trainable parameters per model, optimizer budget, row keys, validation-only checkpoint rule, and split-conformal calibration rule. Three seeds yield 30 trained models and 80,000 identical test rows per seed. The sparse heavy-tail slice contains 139 rows. Table 23 gives the stress test against each control.

Table 23: Frozen probability-law stress test. Positive percentages favor the shared fitting-time candidate. The cluster column is the equal-domain entity bootstrap of candidate minus control CRPS. Every promotion condition fails.

Control	CRPS wins	Cluster difference [95% interval]	Median IS	Slice CRPS	Slice IS	Width failures
shared amortized	4/8	$+8.65$ $[-11.12, 29.03]$	$+0.12\%$	-2.27%	-0.31%	6/8
isolated per domain	3/8	-19.82 $[-52.76, 6.63]$	n/a	-5.53%	-13.64%	5/8

The candidate also misses the 0.765 coverage floor on 30-day funding amount, where coverage is 0.6635 despite mean width near 17,939. Both shared variants select epoch 2 in every seed; isolated fits select epochs 5 through 30, with median 11.5. A read-only decomposition finds that raw predictive width, rather than conformal extension, dominates most width excess. Row-level rank correlation between local-update magnitude and CRPS gain changes sign across domains. Thus the local coefficient is not a stable cross-domain value signal, and this stronger shared successor is not promoted. All six failed conditions, all 30 tasks, and an independent 337-value zero-error recomputation are retained.

E.7 Matched final calibration of frozen fits

The additional comparison reuses every checkpoint from the three-seed, three-fold, four-model study. It fits no new weights. Each held-out series uses its first 40 opened origins for calibration and its following 12 for evaluation, retaining only calibration targets ending by the first evaluation cutoff. All 153 series meet this rule with 40 calibration origins. At each of the 16 normalized horizon coordinates, signed joint calibration uses rank 33 at target coverage 80 percent; crossed final endpoints collapse to their midpoint. No evaluation outcome updates this fixed pool.

Evaluation paths use the original generator and row order, before drawing calibration paths with a separately fixed seed. All original raw metrics are reconstructed from the saved checkpoints. Intervals use 64 path draws per origin. Primary inference averages origins within series and then the three seeds, resampling series within each fixed dataset 10,000 times. Per-seed and other metric intervals are descriptive; the single primary contrast is the pooled final interval score against the amortized

control. Energy remains a metric of the original path law because endpoint adjustment does not specify a new joint distribution. The resampled and clipped target is not a native-frequency task, and measurement timestamps are not verified label publication times.

These bootstrap intervals condition on the fitted checkpoints and the fixed datasets. They do not include retraining uncertainty or unknown dependence between different series from the same dataset. In particular, the larger number of origins is not treated as a larger number of independent series. The complete four-model and per-dataset results follow; their different weightings do not replace the primary estimand. Tables 24, 25 and 26 give score, width and coverage before and after matched final calibration by dataset, and Table 27 gives the complete seed-by-model results with equal series weights.

E.8 All frozen models under matched calibration

All four models use the same calibration and evaluation origins. Scores and widths are in the normalized target units of this experiment, not original physical units. The autoregressive and Gaussian controls have their own likelihoods; only the local and amortized pair shares the identical parameterization. Tables below retain every model and dataset. The target coverage is 0.80, not a statement that observed coverage equals 0.80.

E.9 Scope of the reproducible record

The finite constructions retain exact rational risks and selection counts. The chronological comparisons retain all requested cells, exclusions, calibration ranks and emitted endpoints. Classical controls and adaptations of native code are labeled separately. Checkpoint reuse and saved-output reconstruction do not imply that the public package retrains every historical model or redistributes third-party sources without permission.

F Chronological refitting and online feedback

This separately frozen comparison tests local step selection followed by final calibration of the same base model. The sources are Appliances Energy Prediction, ten-minute energy use in one building, and Metro Interstate Traffic Volume, hourly traffic at one station (Candanedo, 2017; Hogue, 2019). Neither source documents label-release times, so availability is an explicit retrospective assumption. It compares known procedures, not a new architecture or an untouched confirmation.

Calendar windows and common inputs. Ten seven-day test windows per source are chosen by equally spaced integer calendar offsets from day 56 to seven days before the source's last complete elapsed day. The requested test weeks do not overlap. Relative to each test start, days $[-28, -7)$ fit all coefficients, $[-7, -5)$ provide internal calibration, $[-5, -3)$ select a step, and $[-3, 0)$ supply final calibration. Every eligible window refits the predictors with one fixed penalty, prior, normalization and feature set, chosen before this comparison. Deterministic fits have no seed replication. Past evaluation targets may become later lagged inputs, but cannot refit coefficients or select steps within that window.

To bound the full-contact calculation, all step selectors receive the same 19 internal-calibration and 32 selection origins, evenly spaced by index within their roles independently of targets. Coefficient fitting and final calibration use all eligible rows in their respective roles. This is a limited-feedback comparison, not a claim that the complete optimizer uses every available internal label. The common direction, bounds and tie rule are fixed in advance. Ordinary grid search on the full contact set is the same estimator as complete search; the 201-point grids have a different computational budget. The joint Student posterior learns its nuisance parameters from the same observations as QR without simulator-known noise widths, but it imposes a different function class. The clipped update also adds a nonlinear feature absent from that ordinary QR fit; shared inputs do not imply a matched action class. The expanded controls in Appendix F.2 address this distinction.

Eligibility and uncertainty. The protocol requires at least 100/19/32/30/30 origins in the five roles and an eligible origin on every test day. All ten energy windows qualify. Three traffic windows do

Table 24: Interval score before and after matched final calibration. Dataset rows average seeds and series within each dataset. The equal-dataset row gives each of the nine datasets equal weight and is descriptive; the primary analysis instead weights the 153 series equally.

Dataset	Model	Raw	Final
Electricity	Amortized	4.8955	4.3044
	Autoregressive	5.3476	4.5410
	Gaussian	5.2815	4.3356
	Local fit	4.7000	4.1797
ETTh1	Amortized	5.9421	5.5646
	Autoregressive	5.8219	5.3766
	Gaussian	5.5175	5.0487
	Local fit	5.5114	5.1538
ETTh2	Amortized	7.5294	7.5858
	Autoregressive	7.7001	7.7320
	Gaussian	9.5653	9.5304
	Local fit	7.5006	7.6047
ETTm1	Amortized	4.3103	3.9132
	Autoregressive	4.1021	3.7521
	Gaussian	4.7614	4.1155
	Local fit	4.4606	3.9123
ETTm2	Amortized	3.9079	3.7906
	Autoregressive	3.8685	3.7531
	Gaussian	4.8505	4.8150
	Local fit	3.9782	3.8569
Exchange	Amortized	4.4112	4.2651
	Autoregressive	10.6368	9.2579
	Gaussian	12.1224	11.6359
	Local fit	4.3436	4.2461
Solar	Amortized	15.7638	13.8108
	Autoregressive	16.8529	14.8349
	Gaussian	18.9133	16.6114
	Local fit	14.5104	13.6744
Traffic	Amortized	4.1236	3.6502
	Autoregressive	5.6621	4.7690
	Gaussian	5.0059	3.7418
	Local fit	4.1382	3.7010
Weather	Amortized	8.5173	7.6915
	Autoregressive	7.5855	7.3259
	Gaussian	8.0403	7.9628
	Local fit	7.8595	7.3788
Equal dataset	Amortized	6.6001	6.0640
	Autoregressive	7.5086	6.8158
	Gaussian	8.2287	7.5330
	Local fit	6.3336	5.9675

Table 25: Width before and after matched final calibration. Dataset rows average seeds and series within each dataset. The equal-dataset row gives each of the nine datasets equal weight and is descriptive; the primary analysis instead weights the 153 series equally.

Dataset	Model	Raw	Final
Electricity	Amortized	3.2437	2.9001
	Autoregressive	3.7112	2.9177
	Gaussian	3.2398	2.8468
	Local fit	3.0982	2.7952
ETTh1	Amortized	4.7608	4.5225
	Autoregressive	4.2850	3.6245
	Gaussian	3.9490	3.4474
	Local fit	4.3775	4.1519
ETTh2	Amortized	4.1862	4.2304
	Autoregressive	3.9360	4.2699
	Gaussian	4.4394	3.9480
	Local fit	3.8262	4.0567
ETTm1	Amortized	3.0747	3.1544
	Autoregressive	2.7864	2.8292
	Gaussian	3.8104	2.9877
	Local fit	3.1828	3.1540
ETTm2	Amortized	2.5802	2.7358
	Autoregressive	2.2755	2.7891
	Gaussian	3.8854	3.7823
	Local fit	2.7122	2.7834
Exchange	Amortized	1.8093	2.0338
	Autoregressive	2.8022	3.7126
	Gaussian	4.2897	4.6803
	Local fit	1.8051	1.9718
Solar	Amortized	5.9257	11.7971
	Autoregressive	3.9763	13.0018
	Gaussian	4.9640	14.8372
	Local fit	6.4191	11.6666
Traffic	Amortized	2.6138	2.7068
	Autoregressive	4.7590	3.8473
	Gaussian	3.5820	2.7056
	Local fit	2.5386	2.7193
Weather	Amortized	3.0893	4.6082
	Autoregressive	3.1547	4.5059
	Gaussian	4.1457	4.8297
	Local fit	2.9737	4.2792
Equal dataset	Amortized	3.4760	4.2988
	Autoregressive	3.5207	4.6109
	Gaussian	4.0339	4.8961
	Local fit	3.4370	4.1753

Table 26: Coverage before and after matched final calibration. Dataset rows average seeds and series within each dataset. The equal-dataset row gives each of the nine datasets equal weight and is descriptive; the primary analysis instead weights the 153 series equally.

Dataset	Model	Raw	Final
Electricity	Amortized	0.7693	0.7846
	Autoregressive	0.8213	0.7741
	Gaussian	0.7937	0.7872
	Local fit	0.7666	0.7862
ETTh1	Amortized	0.8264	0.8212
	Autoregressive	0.8110	0.7798
	Gaussian	0.8537	0.7999
	Local fit	0.8326	0.8393
ETTh2	Amortized	0.7889	0.7981
	Autoregressive	0.7644	0.7984
	Gaussian	0.8038	0.7413
	Local fit	0.7679	0.8011
ETTm1	Amortized	0.7934	0.8378
	Autoregressive	0.7830	0.8073
	Gaussian	0.8571	0.7626
	Local fit	0.7733	0.8249
ETTm2	Amortized	0.7641	0.8100
	Autoregressive	0.7331	0.8194
	Gaussian	0.8331	0.8046
	Local fit	0.7753	0.8145
Exchange	Amortized	0.6443	0.7237
	Autoregressive	0.4861	0.6022
	Gaussian	0.5653	0.5805
	Local fit	0.6530	0.7179
Solar	Amortized	0.5368	0.8424
	Autoregressive	0.6220	0.8627
	Gaussian	0.6861	0.8646
	Local fit	0.6465	0.8438
Traffic	Amortized	0.8119	0.8218
	Autoregressive	0.9153	0.8283
	Gaussian	0.9113	0.8130
	Local fit	0.7920	0.8099
Weather	Amortized	0.7217	0.8384
	Autoregressive	0.7445	0.8512
	Gaussian	0.8304	0.8444
	Local fit	0.7216	0.8313
Equal dataset	Amortized	0.7397	0.8087
	Autoregressive	0.7423	0.7915
	Gaussian	0.7927	0.7776
	Local fit	0.7476	0.8076

not: the November 2014 window has no records, the June 2015 window lacks sufficient fitting and internal feedback, and the February 2016 window lacks selection and final-calibration rows. None is replaced. The seven remaining traffic windows are all reported. The evaluation contains 10,080 energy and 1,104 traffic origins, not that many independent trials.

Source-specific means weight days and weeks equally. Final complete-search contrasts against no update and raw-loss grid selection use circular length-two window bootstrap intervals with 9,999 draws and a four-comparison adjustment. These are approximate intervals with only ten energy windows and overlapping fitting histories, not guaranteed confidence coverage. With fewer than eight eligible windows, the frozen rule permits only descriptive differences, as in traffic. The prefix variance estimate does not justify a power guarantee. No source-population interval is inferred from two sites.

Separate online protocol. The online panel uses the unmodified ACI and trailing-window function bodies from the Conformal PID authors' implementation. It does not reproduce the original ACI authors' complete forecasting pipeline (Gibbs & Candès, 2021; Angelopoulos et al., 2023). Signed joint residuals come from the same frozen QR endpoints. The final-calibration observations warm the stream; thereafter each test interval precedes its own label, and past responses update only calibration. ACI uses learning rate 0.005, burn-in 20 and a 256-observation window. Rolling calibration uses that same window. Windows count retained observations rather than elapsed time under gaps. Infinite thresholds are retained, though none occurs in the scored rows. We sort raw endpoints and collapse any inverted final interval to its midpoint; these conventions and the source's quantile implementation are not a general finite-time validity theorem. This QR panel does not reproduce SPCI, bias-corrected ACI, RetroAdj or a learned online forecaster.

Table 28 gives the chronological comparison at nominal coverage 0.9, weighting days and then test weeks equally.

Outcome. Complete and ordinary same-contact searches emit identical final endpoints in all 17 eligible windows. Complete search worsens mean final score against ordinary quantiles by 4.76 on energy and 290.83 on traffic. The approximate energy interval is $[-0.42, 12.20]$; improvement occurs in 3/10 energy and 1/7 traffic windows. Against raw-loss grid selection, final differences are 2.27 and 94.46; the energy interval is $[-3.33, 9.78]$. These results do not establish a useful selected update. The raw comparison already favors ordinary QR, so the result cannot be described solely as calibration erasing a beneficial fit. Online feedback reduces mean score slightly on both sources, but its coverage change differs by source and its information rights differ from fixed calibration. It is not evidence that one common procedure dominates all criteria.

Reconstructed losses and internal objectives differ by at most 2.92×10^{-11}. Coverage uses original-unit endpoints without tolerance. Unit conversion changes eight seasonal coverage summaries by 1–6 boundary observations each; original and recomputed summaries remain archived. Models and endpoints are unchanged.

F.1 Native bootstrap and sequential residual fitting

A later development protocol runs the SPCI authors' bootstrap fitter and sequential quantile-forest implementation (Xu & Xie, 2023), rather than injecting centers from a different forecaster. The supported linear-regression configuration uses its default weighted least squares, with decay 0.99, 50 bootstrap fits, unit residual scale and one fixed random seed per window. We retain the authors' weighting by bootstrap sample position; it is not weighting by original calendar age. All eligible observations in the preceding 28 days fit the model. Features match the earlier comparison, with training-only normalization. SPCI uses 100 lagged residuals, ten trees of depth two and five tail-allocation bins. EnbPI uses the identical bootstrap predictions and out-of-bag residuals, replacing conditional quantile prediction with the authors' rolling empirical quantiles. Neither method refits the point model inside the test week. The static residual reference uses fixed empirical training-residual quantiles around that same point prediction.

The same twenty historical windows yield seventeen eligible windows, all reported in Table 29. The three previous exclusions are not replaced. Every current interval precedes its response in the online calibrator; earlier test labels may update later intervals. Synthetic tests mutating current or future

responses and future features preserve previously issued prefixes. Measurement timestamps still stand in for label availability by assumption, not observed release records. This is a historical replay on already opened data, not a prospective confirmation. Native empirical quantile extremes are retained without a finite-time coverage claim. No training row lacks an out-of-bag predictor in the executed fits.

SPCI has lower descriptive score than its shared-fit EnbPI control in both sources. Its comparison with earlier QR is mixed: energy has lower width and higher coverage but worse score; traffic has lower score and width with higher coverage. The earlier complete search does not acquire a superiority claim from either comparison. These are source-specific observations without new significance intervals or a source-population guarantee. One native seed per window does not quantify stochastic fitting variation, and overlapping training histories remain dependent.

All 34 native method-window outputs and the unchanged comparison rows are archived. Reconstructing native metrics from the saved endpoints gives zero discrepancy. Bootstrap fitting, prediction and out-of-bag construction total 7.98 seconds; sequential SPCI quantile refitting totals 4,413.18 seconds and EnbPI interval construction 13.61 seconds across the eligible windows. These phase totals exclude imports, source loading and result serialization. They are not full wall-clock or equal-budget superiority claims. Dependency bridges cover removed import symbols, renamed criterion/API arguments and single-query array shape, not a change to the authors' fitting or interval formulas. The published electricity benchmark and its tuning are not reproduced. Bias-corrected ACI and RetroAdj remain absent from this real-data comparison. A subsequent RetroAdj check runs the unchanged author R code with its supported single-rate ACI controller: 32 initial observations and 32 synthetic queries pass own-outcome and future-feature prefix checks. This establishes executability under immediate feedback, not delayed-feedback parity, real-domain effectiveness or a completed comparator benchmark.

F.2 Matching the clipped feature and delaying feedback

The scalar update uses $g(x) = \mathrm{clip}((x_{\mathrm{recent}} - x_{\mathrm{day}})/\hat{\sigma}, -3, 3)$. This function need not lie in the original linear feature span. Sharing input fields therefore does not match the model class. We add an ordinary quantile regressor containing this same feature, with an independently learned coefficient for each tail. It jointly fits the intercept, calendar terms and lag coefficients, using the same 0.001 penalty on non-intercept coefficients. It has 16 fitted coefficients across the two tails, versus 14 without the clipped feature. Another control fits those 16 coefficients on the fitting role plus the same 19 internal-calibration and 32 selection responses available to the scalar selectors. A joint Student posterior uses the same augmented design and pooled labels under the previous fixed priors. This is an ordinary linear posterior, not a learned recurrent or meta-learned adapter.

The scalar raw-loss and calibrated-loss selectors retain the same base fit, shared center direction, 201 steps in $[-1, 1]$ and tie rule. They are the matched action-family comparison. Joint two-tail QR is more flexible and penalizes its additional coefficients; the scalar selection objective does not penalize its step. Complete and same-event search remain the identical estimator by definition. No configuration or seed search is performed.

One release rule for all methods. We retain the twenty calendar episodes and cross each with delays of zero and one native tick. The source timestamp is treated as the right endpoint of its measurement interval, with issuance one tick earlier. Both this interval convention and release at the timestamp plus the assumed delay are retrospective assumptions, not verified publication records. At an issue, all responses whose release time is no later than that issue become available before forecasting. Every model's recent and daily lag shifts under the same rule. Exact required history must exist; an adjacent stored row cannot bridge a missing calendar interval.

The same availability rule restricts training labels before internal calibration, selection labels before final calibration, and final-calibration labels before testing. The ordinary native bootstrap uses all mature pre-test roles. A pre-test response excluded by a role cutoff is not subsequently inserted into that initial history. During testing, only feedback for earlier issued eligible predictions enters online residuals and controller updates. Methods reset at each episode. Residual lags count this retained sequence, not every native calendar tick. These explicit restrictions do not identify the unknown true deployment delay or guarantee robustness to missing data.

The online panel receives no extra split calibration. ACI and rolling quantiles follow the cited author formulas under the release scheduler. ACI updates its error level using the residual-threshold exceedance only when the corresponding response arrives. SPCI retains the author's 50-bootstrap linear fits and residual quantile-forest routines; its requested retained-sequence horizon counts prior issued but unreleased responses. EnbPI uses the same centers and released residuals with the author's empirical-quantile rule. Synthetic zero-delay fixtures reproduce the original loops, while delayed-response and future-feature mutations leave already issued prefixes unchanged. The release scheduler is an adaptation, not an unchanged native delayed- feedback implementation or a new coverage theorem.

Residual-threshold feedback is not generally identical to coverage of a midpoint-collapsed interval: raw $[-1, 1]$, correction -2 and outcome 0 give a residual exceedance but a covered emitted point. A read-only reconstruction checks every retained ACI trajectory, including warmup: 31,946 rows contain 30 crossings but no feedback discrepancy. Replaying the same fitted predictions with actual-emission feedback changes zero of the 22,363 test intervals, with zero reconstruction error. Thus this boundary case does not change the published results; it remains an implementation limitation outside the checked trajectories. No models were refitted.

Denominators and inference. All methods emit on the same eligible origins within a delay. We separately report each delay's available origins and the cross-delay timestamp intersection fixed by history eligibility before fitting. Neither set is selected from method performance. Training and test requirements, absent days, failed cells and the unchanged episode schedule remain explicit. Source means weight days and then complete seven-day episodes equally.

The primary family has eight final-score contrasts: calibrated-loss selection minus raw-loss selection or joint clipped QR, for two sources and two delays on common origins. It uses the predetermined circular episode bootstrap with 9,999 draws, length two and a two-sided Bonferroni level of 99.375 percent. Lengths one and three are reported as sensitivity checks, not selected from their results. Fewer than eight complete episodes receive no interval. Sparse episodes and overlapping fitting histories do not ensure nominal coverage or power; raw rows are not independent repetitions. Other method, raw-score and available-origin comparisons are descriptive. Complete results are in Table 31; Figure 2 compares the same fixed predictors before and after final calibration.

Results. All forty requested episode–delay cells terminate. Thirty-four qualify, giving 408 fitted method cells. The same three traffic episodes are ineligible under both delays; none is replaced. Each energy delay has ten episodes and 10,080 test origins. Traffic has seven episodes, 1,104 available origins without extra delay and 1,099 with it; the predetermined common set has 1,065 origins. Days are consecutive 24-hour bins anchored to the episode start, not necessarily midnight-to-midnight dates.

The clipped feature is active in energy: it clips 1,592/30,240 and 1,593/30,230 fitting rows under the two delays, respectively. It clips 548/10,080 and 545/10,080 available test rows. These counts sum fitting-role occurrences across overlapping episodes, not distinct observations. No traffic fitting or test row is clipped. Even without clipping, redundant features can change an L1-penalized fit because they change the coefficient representation of the penalty. Sharing a feature span alone therefore does not establish equality of these fitted estimators.

Calibrated-loss selection has higher final score than joint clipped QR in all four settings. With delayed energy labels, its difference is $+39.27$ and the primary adjusted interval is $[5.39, 94.47]$; all three block lengths are unfavorable. The other energy primary intervals include zero, not evidence of equivalence. On delayed traffic, calibration-aware selection improves descriptively over raw-loss selection by 299.93, in six of seven episodes, but remains worse than joint QR by 272.46. Thus a favorable objective contrast in one setting does not establish a competitive model. Table 30 preserves all eight primary contrasts and every prespecified block sensitivity.

The online panel further separates feedback value from model comparison. The rolling clipped-QR score is $230.35/307.78$ on energy, versus $232.41/309.64$ for fixed joint QR. On traffic, SPCI changes from 1943.60 without extra delay to 3895.26 with it, compared with $2214.82/3329.81$ for joint QR. These are descriptive comparisons with different training-label and feedback budgets, not controlled attribution to one algorithm component. SPCI remains lower in score than its shared-fit EnbPI control

in all four settings. Coverage and both tail errors are reported alongside score and width; score alone is not a coverage guarantee.

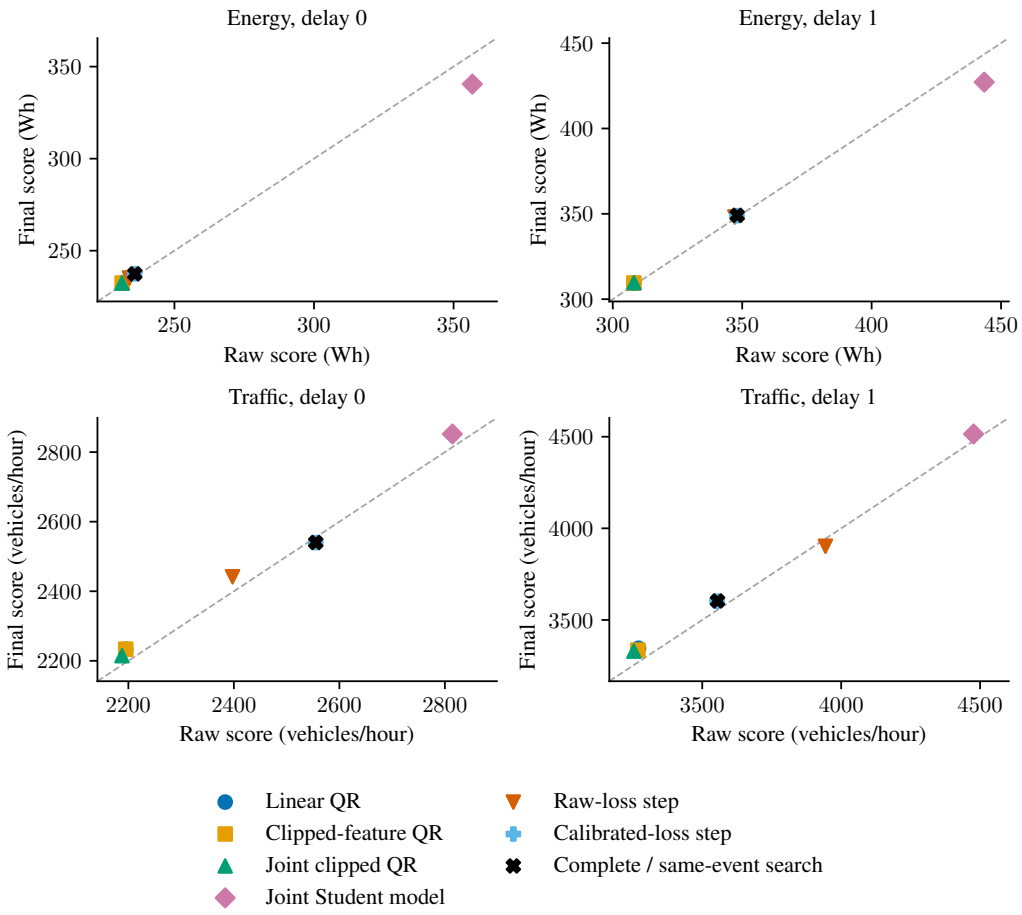

Figure 2: Raw and final interval scores on predetermined common origins, with equal-day then equal-episode weighting. Lower scores are better. Points below the dashed identity line improve after calibration; this is not superiority over another predictor or evidence identifying the cause of an individual failure. Complete search and the same-event reference share one legend entry because they coincide, without artificial jitter. Panels use native target units and explicit hypothetical delays; they are descriptive means, not independent-domain confidence intervals.

Table 27: Complete seed-by-model results with equal series weights. Energy is evaluated only on the raw path law; no final joint law is defined by endpoint correction.

Seed	Model	Raw IS	Final IS	Final width	Final cov.	Energy
1	Amortized	7.6214	6.9404	5.1621	0.8152	1.2702
	Autoregressive	8.2842	7.3926	5.5468	0.8142	1.3381
	Gaussian	9.0624	8.0172	5.8975	0.8030	1.4304
	Local fit	7.2763	6.7486	4.9928	0.8150	1.2418
2	Amortized	7.4348	6.6307	4.8201	0.8122	1.2493
	Autoregressive	8.5759	7.6230	5.6718	0.8085	1.3654
	Gaussian	8.9820	7.9156	5.7946	0.8053	1.4409
	Local fit	7.0808	6.5850	4.7599	0.8089	1.1872
3	Amortized	7.6700	6.7841	5.1644	0.8163	1.2542
	Autoregressive	8.3642	7.4322	5.4765	0.8089	1.3364
	Gaussian	8.8793	7.8891	5.7685	0.8037	1.4335
	Local fit	7.1548	6.7131	5.0405	0.8121	1.2106

Table 28: Chronological comparison at nominal coverage 0.9. Means weight days and then test weeks equally, in source-specific units. The final two rows in each source use online feedback, unlike the fixed-calibration rows.

Method	Raw IS	Final IS	Coverage	Width
Appliances: 10 evaluable test weeks				
Seasonal residual	677.27	727.58	0.8742	339.15
Linear quantiles	231.52	232.65	0.8866	123.53
Joint Student posterior	357.01	340.53	0.8893	120.89
Raw-loss grid step	233.89	235.14	0.8865	126.06
Calibrated-loss grid step	235.69	237.40	0.8889	125.80
Complete contact search	235.80	237.42	0.8899	125.94
Ordinary same-contact grid	235.80	237.42	0.8899	125.94
QR + author-source ACI	231.52	230.74	0.8999	124.44
QR + rolling calibration	231.52	230.56	0.8976	123.43
Metro traffic: 7 evaluable test weeks				
Seasonal residual	5399.40	5621.54	0.8783	2663.33
Linear quantiles	2226.12	2264.20	0.9089	1821.64
Joint Student posterior	2867.56	2913.95	0.8879	1850.66
Raw-loss grid step	2423.66	2460.57	0.8971	1869.46
Calibrated-loss grid step	2582.79	2554.91	0.8913	2007.69
Complete contact search	2583.12	2555.03	0.8913	2005.01
Ordinary same-contact grid	2583.12	2555.03	0.8913	2005.01
QR + author-source ACI	2226.12	2244.33	0.8875	1757.92
QR + rolling calibration	2226.12	2238.66	0.8900	1751.62

Table 29: Native bootstrap comparison at nominal coverage 90 percent. All means weight days then eligible weeks equally within each source. Score and width use native target units; coverage and tail misses are fractions. Earlier QR and complete search are unchanged contextual results, not matched-training or matched-feedback controls. For traffic these two rows are not the same quantity as in Table 40, whose rows use the 1,065 origins common to all twelve methods.

Source	Method	Weeks	Score	Coverage	Width	Left miss	Right miss
Energy	Static residual	10	329.29	0.9035	122.31	0.0458	0.0507
Energy	SPCI	10	281.67	0.9085	109.37	0.0265	0.0650
Energy	EnbPI	10	342.00	0.9003	111.35	0.0299	0.0698
Energy	Earlier QR	10	232.65	0.8866	123.53	0.0614	0.0520
Energy	Earlier complete search	10	237.42	0.8899	125.94	0.0590	0.0511
Traffic	Static residual	7	2810.09	0.9022	1958.82	0.0510	0.0468
Traffic	SPCI	7	2109.45	0.9197	1611.00	0.0383	0.0421
Traffic	EnbPI	7	2875.00	0.8976	1887.38	0.0278	0.0746
Traffic	Earlier QR	7	2264.20	0.9089	1821.64	0.0471	0.0439
Traffic	Earlier complete search	7	2555.03	0.8913	2005.01	0.0568	0.0519

Table 30: The complete primary family. Differences are final interval score of calibrated-loss selection minus the named control, so negative favors selection. Raw denotes raw-loss selection; QR denotes joint clipped QR. Delay is in native ticks. Each interval uses the eight-member Bonferroni level of 99.375 percent. Block length two is primary; lengths one and three are sensitivity checks. Traffic has fewer than eight episodes and receives no interval. All intervals are approximate, without a power guarantee.

Source, delay	Control	Difference	Block 1	Block 2	Block 3
Energy, 0	Raw	$+2.26$	$[-3.88, 9.77]$	$[-3.99, 10.00]$	$[-2.97, 9.13]$
Energy, 0	QR	$+4.99$	$[0.05, 12.08]$	$[-0.38, 12.60]$	$[-0.26, 11.91]$
Energy, 1	Raw	$+0.79$	$[-24.89, 16.58]$	$[-27.43, 18.12]$	$[-22.12, 17.98]$
Energy, 1	QR	$+39.27$	$[3.89, 89.21]$	$[5.39, 94.47]$	$[6.29, 85.27]$
Traffic, 0	Raw	$+98.36$	n/a	n/a	n/a
Traffic, 0	QR	$+325.05$	n/a	n/a	n/a
Traffic, 1	Raw	-299.93	n/a	n/a	n/a
Traffic, 1	QR	$+272.46$	n/a	n/a	n/a

Table 31: Shared-release comparison on predetermined common origins. Means weight days then complete episodes equally. Score and width use source-specific units. Native online rows have different fitting and feedback rights from fixed-calibration rows.

Method	Stage	Weeks	Score	Coverage	Width	Left	Right
Energy; assumed delay 0 native tick							
Linear QR	Raw	10	231.52	0.9119	127.70	0.0422	0.0459
Linear QR	Final	10	232.65	0.8866	123.53	0.0614	0.0520
Clipped-feature QR	Raw	10	231.28	0.9120	127.12	0.0419	0.0461
Clipped-feature QR	Final	10	232.40	0.8871	123.11	0.0607	0.0522
Joint clipped QR	Raw	10	231.21	0.9129	126.58	0.0408	0.0463
Joint clipped QR	Final	10	232.41	0.8865	122.35	0.0613	0.0522
Joint Student model	Raw	10	356.72	0.9404	214.41	0.0198	0.0398
Joint Student model	Final	10	340.50	0.8938	122.44	0.0477	0.0584
Raw-loss step	Raw	10	233.89	0.9036	127.70	0.0502	0.0462
Raw-loss step	Final	10	235.14	0.8865	126.06	0.0636	0.0499
Calibrated-loss step	Raw	10	235.69	0.9064	127.70	0.0459	0.0476
Calibrated-loss step	Final	10	237.40	0.8889	125.80	0.0598	0.0513
Complete search	Raw	10	235.80	0.9063	127.70	0.0461	0.0475
Complete search	Final	10	237.42	0.8899	125.94	0.0590	0.0511
Same-event reference	Raw	10	235.80	0.9063	127.70	0.0461	0.0475
Same-event reference	Final	10	237.42	0.8899	125.94	0.0590	0.0511
Clipped QR + ACI	Final	10	230.51	0.9000	124.03	0.0477	0.0523
Clipped QR + rolling	Final	10	230.35	0.8974	122.92	0.0495	0.0531
Bootstrap + SPCI	Final	10	285.25	0.9084	109.92	0.0282	0.0634
Bootstrap + EnbPI	Final	10	341.65	0.9011	112.03	0.0309	0.0681
Energy; assumed delay 1 native tick							
Linear QR	Raw	10	308.14	0.9094	179.90	0.0412	0.0494
Linear QR	Final	10	309.44	0.8722	172.89	0.0688	0.0590
Clipped-feature QR	Raw	10	308.14	0.9090	179.12	0.0410	0.0500
Clipped-feature QR	Final	10	309.50	0.8723	172.21	0.0678	0.0599
Joint clipped QR	Raw	10	308.16	0.9086	179.86	0.0414	0.0500
Joint clipped QR	Final	10	309.64	0.8710	172.70	0.0688	0.0602
Joint Student model	Raw	10	443.50	0.9404	267.40	0.0130	0.0466
Joint Student model	Final	10	427.16	0.8878	157.95	0.0468	0.0654
Raw-loss step	Raw	10	347.15	0.8844	179.90	0.0588	0.0567
Raw-loss step	Final	10	348.13	0.8907	194.20	0.0576	0.0517
Calibrated-loss step	Raw	10	347.94	0.8783	179.90	0.0644	0.0573
Calibrated-loss step	Final	10	348.91	0.8939	197.75	0.0548	0.0513
Complete search	Raw	10	348.08	0.8781	179.90	0.0646	0.0573
Complete search	Final	10	349.11	0.8941	198.12	0.0547	0.0512
Same-event reference	Raw	10	348.08	0.8781	179.90	0.0646	0.0573
Same-event reference	Final	10	349.11	0.8941	198.12	0.0547	0.0512
Clipped QR + ACI	Final	10	309.53	0.8993	179.30	0.0470	0.0537
Clipped QR + rolling	Final	10	307.78	0.8961	175.99	0.0498	0.0541
Bootstrap + SPCI	Final	10	397.21	0.9063	167.57	0.0260	0.0677
Bootstrap + EnbPI	Final	10	431.78	0.9010	156.74	0.0284	0.0706

Method	Stage	Weeks	Score	Coverage	Width	Left	Right
Traffic; assumed delay 0 native tick							
Linear QR	Raw	7	2194.98	0.8990	1738.33	0.0482	0.0528
Linear QR	Final	7	2233.45	0.9112	1818.05	0.0460	0.0429
Clipped-feature QR	Raw	7	2194.98	0.8990	1738.33	0.0482	0.0528
Clipped-feature QR	Final	7	2233.45	0.9112	1818.05	0.0460	0.0429
Joint clipped QR	Raw	7	2187.88	0.9031	1748.38	0.0493	0.0476
Joint clipped QR	Final	7	2214.82	0.9098	1803.59	0.0464	0.0438
Joint Student model	Raw	7	2814.20	0.9151	2004.76	0.0238	0.0611
Joint Student model	Final	7	2852.08	0.8902	1849.62	0.0394	0.0704
Raw-loss step	Raw	7	2397.45	0.8878	1738.33	0.0620	0.0502
Raw-loss step	Final	7	2441.52	0.8965	1865.88	0.0593	0.0442
Calibrated-loss step	Raw	7	2555.33	0.8526	1738.33	0.0828	0.0646
Calibrated-loss step	Final	7	2539.87	0.8919	2004.11	0.0579	0.0503
Complete search	Raw	7	2555.22	0.8526	1738.33	0.0828	0.0646
Complete search	Final	7	2540.21	0.8919	2001.43	0.0579	0.0503
Same-event reference	Raw	7	2555.22	0.8526	1738.33	0.0828	0.0646
Same-event reference	Final	7	2540.21	0.8919	2001.43	0.0579	0.0503
Clipped QR + ACI	Final	7	2217.14	0.8893	1754.25	0.0573	0.0535
Clipped QR + rolling	Final	7	2209.83	0.8917	1747.98	0.0539	0.0543
Bootstrap + SPCI	Final	7	1943.60	0.9299	1611.83	0.0342	0.0358
Bootstrap + EnbPI	Final	7	2825.18	0.9008	1887.41	0.0271	0.0720
Traffic; assumed delay 1 native tick							
Linear QR	Raw	7	3270.74	0.8948	2774.79	0.0502	0.0550
Linear QR	Final	7	3345.53	0.9038	2878.81	0.0455	0.0507
Clipped-feature QR	Raw	7	3268.30	0.8929	2766.28	0.0502	0.0569
Clipped-feature QR	Final	7	3335.31	0.9042	2874.96	0.0434	0.0525
Joint clipped QR	Raw	7	3253.86	0.8976	2795.80	0.0430	0.0594
Joint clipped QR	Final	7	3329.81	0.9095	2925.34	0.0428	0.0477
Joint Student model	Raw	7	4476.46	0.9272	3103.27	0.0027	0.0700
Joint Student model	Final	7	4514.86	0.8878	2731.97	0.0229	0.0893
Raw-loss step	Raw	7	3943.33	0.8532	2774.79	0.0746	0.0722
Raw-loss step	Final	7	3902.20	0.9025	3195.67	0.0450	0.0525
Calibrated-loss step	Raw	7	3553.05	0.8653	2774.79	0.0698	0.0649
Calibrated-loss step	Final	7	3602.27	0.9164	3106.85	0.0362	0.0474
Complete search	Raw	7	3555.02	0.8653	2774.79	0.0698	0.0649
Complete search	Final	7	3603.66	0.9139	3107.04	0.0379	0.0482
Same-event reference	Raw	7	3555.02	0.8653	2774.79	0.0698	0.0649
Same-event reference	Final	7	3603.66	0.9139	3107.04	0.0379	0.0482
Clipped QR + ACI	Final	7	3301.91	0.8929	2801.73	0.0477	0.0594
Clipped QR + rolling	Final	7	3303.33	0.8954	2802.83	0.0479	0.0567
Bootstrap + SPCI	Final	7	3895.26	0.8939	2644.03	0.0354	0.0707
Bootstrap + EnbPI	Final	7	4262.43	0.9042	2772.92	0.0248	0.0710

F.3 Native retrospective adjustment with bounded memory

A separate comparison runs the unmodified author implementation of RetroAdj (Jun & Ohn, 2026) with its ACI and DtACI (Gibbs & Candès, 2024) branches. Each model uses the latest 168 mature observations, an RBF kernel, regularization 0.1, and the author's training-only bandwidth rule. The single-gain branch uses 0.005; DtACI uses the author's eight-gain default grid. No validation tuning or source modification is performed. The source wrapper does not forward its optional bandwidth argument to KRR, so this comparison retains the implemented default rather than claiming a manually specified bandwidth. The algorithm's known sequence length is the fixed retrospective episode length, not future label values.

Both native branches and ordinary rolling QR receive the same standardized lag values, calendar features and clipped lag contrast. Normalization uses only the initial 168 mature labels. QR refits its two tails at levels $0.05, 0.95$, with intercept and L1 penalty 0.001, using the oldest 149 of the last 168 observations. The remaining 19 give a nonnegative joint rank-18 correction. RetroAdj instead uses all 168 model rows and its native jackknife adjustment. Thus available history is matched, but model classes, fitting allocation and computational cost are not.

At each issue, earlier feedback is processed before predicting the next target. Native source tests change own/future labels and future covariates without changing the corresponding issued prefix. This experiment assumes immediate release at the target interval's right endpoint; it does not establish actual publication times or a native delayed-feedback comparison. It does not refit any method in the earlier full-history SPCI/EnbPI panel.

All 20 requested windows have terminal records. Seventeen qualify, yielding 51 method cells; three traffic exclusions remain unreplaced. Energy has 10,080 test rows, traffic 1,104. Means weight rows within days, seven days within episodes, then episodes equally. All episode differences and emitted endpoints are retained; no significance or power claim is attached.

Source	Method	Episodes	Score	Coverage	Width
Energy	RetroAdj (ACI)	10	357.36	0.9067	174.51
Energy	RetroAdj (DtACI)	10	304.59	0.9073	180.96
Energy	Rolling joint QR	10	287.27	0.9005	190.06
Traffic	RetroAdj (ACI)	7	1743.19	0.9206	1175.51
Traffic	RetroAdj (DtACI)	7	1715.86	0.9101	1149.68
Traffic	Rolling joint QR	7	2892.59	0.9323	2487.98

Energy favors rolling QR in mean score: 287.27 versus 357.36 and 304.59 for the two native branches. Traffic favors RetroAdj: 1743.19 and 1715.86 versus 2892.59. Mean coverage ranges from 0.9005 to 0.9323 across these six source–method combinations. The result is source-dependent, not a claim that one native method dominates the field. The cited author repository has no identified redistribution license; the artifact records its pinned acquisition instructions rather than redistributing that code.

G AVERAGING OVER FINITE CALIBRATION SAMPLES

G.1 RISK TARGET AND EXACT REFERENCE

Let $1 \le m < N$. Fix an upstream predictor family independently of N IID mature observations D_N. For an action t, let $R_m(t) = \mathbb{E}_{C_m,Z} S_\alpha(I_{t,C_m}(X), Y)$, where C_m and $Z = (X, Y)$ are fresh independent samples. The emission rule, including any nonnegative correction or midpoint collapse, is part of I. Minimizing a loss conditional on one observed calibration pool is not the same as minimizing this expectation.

For each query j, average its loss over all m-element calibration subsets of the remaining $N-1$ observations. For a joint residual rank $k = \lceil (m+1)(1-\alpha) \rceil \le m$, order those $N-1$ residuals, breaking ties by index. The probability that the selected kth rank occupies position r is

$$w_r = \frac{\binom{r-1}{k-1}\binom{N-1-r}{m-k}}{\binom{N-1}{m}}.$$

To prove the identity, choose $k-1$ indices below r, include r, then choose $m-k$ above it. These disjoint events partition all subsets. Equal residuals yield equal corrections, so stable tie-breaking does not affect the averaged loss. Apply the signed correction, optional nonnegative truncation and actual emission rule after each rank. If $k > m$, the interval and its score are infinite; an all-zero collection of binomial weights is not a probability distribution. Separate-tail ranks depend on the same subset and cannot generally be averaged by multiplying their marginal rank weights. The reference comparison below uses joint residuals only.

Assume integrable emitted loss for the expectation statement and square integrable emitted loss for the variance statement. These conditions exclude an overflowing infinite interval from the finite-variance claim. For fixed t, the resulting estimator $\widehat{R}_m(t)$ is unbiased for $R_m(t)$: every disjoint calibration/query tuple is IID, and expectation commutes with the finite average. It is also the conditional expectation of the single-pool loss under a uniformly random permutation of D_N, where the first m observations calibrate and the remaining observations are queries. The conditional variance identity therefore gives

$$\mathrm{Var}(\widehat{R}_m(t)) \le \mathrm{Var}(\widehat{R}_m^{\mathrm{single}}(t)).$$

These are classical cross-validation and conditional-averaging facts (Haberman, 2019); they do not prove that minimizing the averaged loss produces a lower-risk action. Nor do they apply unchanged when the candidate is fitted on the same query labels. Joint regression controls below use fresh final calibration and evaluation, without claiming unbiased internal risk estimates for their fitted coefficients.

G.2 LOWER ESTIMATOR VARIANCE, HIGHER SELECTED RISK

Let X and Y be independent and uniform on $\{-1, 1\}$, $N = 3$, $m = 1$ and $\alpha = 1/2$. Consider all nine raw interval candidates obtained by sorting $[-1 + aX, 1 + bX]$, for $(a, b) \in \{-1, 0, 1\}^2$. Use nonnegative joint calibration. Both selectors minimize their respective exact loss, breaking ties by $a^2 + b^2$, then lexicographically. The single-pool selector uses the first observation for calibration and the other two for selection. The averaged selector uses all six ordered calibration/query pairs.

There are 64 equally likely mature samples. The following table follows by substituting the four possible (X, Y) pairs; final risk averages all 16 fresh calibration/query pairs. Each of the four one-tail actions has risk $13/4$; keep has risk 2. The remaining four actions are never selected.

Action	Final risk	Single-pool count	Averaged count
$(0, 0)$	2	52	36
Each of $(0, \pm 1), (\pm 1, 0)$	$13/4$	3	7
Each of $(\pm 1, \pm 1)$	4	0	0

Thus the selected expected risk increases from $143/64$ to $163/64$, a difference of $5/16$. Meanwhile the fixed-action variance decreases by $5/24$ for each one-tail action and by $1/3$ for each two-tail action; it is unchanged for keep. Since both keep estimates are identically 2, the variance of each

paired loss difference from keep decreases by exactly the same amount. The failure therefore cannot be repaired simply by replacing individual variance with paired-difference variance.

The following four-state calculation explains the selection counts for the single-tail action $(0, 1)$; symmetry covers the other three actions.

State	X	Y	Score at correction 0	Score at correction 1
A	-1	-1	1	3
B	-1	$+1$	5	3
C	$+1$	-1	3	5
D	$+1$	$+1$	3	5

Only B requires a positive correction. With one calibration observation, single-pool selection favors this action precisely for AAA, CAA and DAA, where the first letter denotes calibration. Averaging instead favors AAA and every permutation of AAC and AAD: seven samples rather than three. Its true excess risk remains $5/4$. An independent direct enumeration reproduces the counts and risks without the rank-weight implementation.

How much the tie rule decides. The single-pool objective attains its minimum at more than one action in 44 of the 64 mature samples; the averaged objective is tied in 12. The declared rule prefers the smaller a^2+b^2 and therefore the unchanged interval, which is what an implementation that defaults to leaving the interval alone would do. Replacing it with uniform-random tie-breaking reverses the comparison: selected risk becomes $271/96$ for the single pool against $175/64$ for averaging, and harmful-selection probability $7/12$ against $17/32$. Both computations use the same enumeration and the same final risks. We report the reversal because it fixes what the construction does and does not establish. It establishes the negative statement, that lower estimator variance does not imply lower selected risk, since a deterministic and natural convention exists under which averaging is worse. It does not establish that subset averaging is the worse selector in general, and no claim in this paper uses it that way. This is a counterexample to a universal selected-risk improvement, not evidence that subset averaging never helps. The maximum observed difference within the specified finite family was used to expose the counterexample, not to define a favorable generalization claim.

G.3 Complete bounded comparison

Before execution, we fixed eight artificial distributions, sample budgets $(N, m) = (2, 1), (3, 1), (5, 3)$, signed and nonnegative joint calibration, and center or two-tail action families. These give 78 configurations. Every ordered mature sample and fresh calibration/query outcome is included. Controls are no update, raw-loss selection on all N labels or only the query role, single-pool calibrated selection, full subset averaging, ordinary joint two-tail regression and a linear Normal–Inverse-Gamma posterior. All have the same N available mature labels, but allocate them differently. Regression fits four coefficients jointly at the two feature values; the restricted action families are not equally flexible. The posterior uses a zero mean, identity precision, shape 2 and scale 1, with no true noise or nuisance parameters. Its Student-t endpoints are numerical; population averages are exact conditional on those computed endpoints.

Compared with single-pool selection, averaging lowers risk in 32 configurations, ties in 29 and raises it in 17. Against keep the counts are 39, 9 and 30; against ordinary joint QR they are 48, 0 and 30. These counts describe specified artificial populations, not a sample of independent domains. The original signed-radius risks $13/4$ and $29/8$ and the centered risks 3 versus 2 are recovered within the same reference. Table 32 gives every primary contrast; all seven methods' scores, coverage and width, exact fractions and selection frequencies are retained in the accompanying data.

The counterexample rules out a variance-reduction-only justification for updating. Theorem 4 gives the conditions that do matter: the pairwise reversal probabilities against a reference and the risk gaps, not the variance of each candidate (Appendix G.6). Arbitrary IID subsets are also not a valid substitute for a temporal sampling and dependence model. No new neural architecture, native-method superiority or real-domain confirmation follows from this reference experiment.

Table 32: Complete finite-support comparison. N mature selection labels and m fresh calibration labels; S denotes signed joint calibration with midpoint collapse, C its nonnegative counterpart. Entries are exact expected final interval-score differences displayed to four decimals: subset averaging minus single-pool selection, no update, or joint two-tail QR. Negative favors averaging. NIG results and exact fractions are in the accompanying data. These are specified artificial populations, not independent benchmark repetitions.

Case	Family	N	m	Rule	Δ single	Δ keep	Δ QR
No sign	center	2	1	S	+0.0000	+1.0000	-0.5000
No sign	two tail	2	1	S	-0.3438	+1.0000	-0.5000
No sign	center	2	1	C	+0.0000	+0.0000	-1.5000
No sign	two tail	2	1	C	-0.6250	+0.3125	-1.1875
No sign	center	3	1	S	+0.0000	+0.5000	-0.9375
No sign	two tail	3	1	S	+0.1719	+0.7578	-0.6797
No sign	center	3	1	C	+0.0000	+0.0000	-1.4375
No sign	two tail	3	1	C	+0.3125	+0.5469	-0.8906
No sign	center	5	3	S	+0.2500	+0.7500	-0.3540
No sign	two tail	5	3	S	+0.1772	+0.7833	-0.3207
No sign	center	5	3	C	+0.0000	+0.0000	-1.1040
No sign	two tail	5	3	C	+0.0090	+0.2529	-0.8511
Weak sign	center	2	1	S	+0.0000	+0.0000	-1.0000
Weak sign	two tail	2	1	S	-0.0625	+0.4062	-0.5938
Weak sign	center	2	1	C	+0.0000	+0.0000	-0.9375
Weak sign	two tail	2	1	C	-0.3125	+0.1562	-0.7812
Weak sign	center	3	1	S	+0.0000	+0.0000	-0.7500
Weak sign	two tail	3	1	S	-0.0469	+0.2734	-0.4766
Weak sign	center	3	1	C	+0.0000	+0.0000	-0.7031
Weak sign	two tail	3	1	C	+0.1562	+0.2734	-0.4297
Weak sign	center	5	3	S	+0.0000	+0.0000	-0.1626
Weak sign	two tail	5	3	S	-0.1594	+0.1393	-0.0233
Weak sign	center	5	3	C	+0.0000	+0.0000	-0.1565
Weak sign	two tail	5	3	C	+0.0045	+0.1265	-0.0300
Strong sign	center	2	1	S	+0.0000	-1.5000	-2.0000
Strong sign	two tail	2	1	S	+0.0312	-1.1562	-1.6562
Strong sign	center	2	1	C	+0.0000	-1.5000	-2.0000
Strong sign	two tail	2	1	C	-0.0625	-1.1250	-1.6250
Strong sign	center	3	1	S	+0.0000	-1.7500	-1.7500
Strong sign	two tail	3	1	S	-0.1016	-1.5156	-1.5156
Strong sign	center	3	1	C	+0.0000	-1.7500	-1.7500
Strong sign	two tail	3	1	C	-0.1250	-1.4531	-1.4531
Strong sign	center	5	3	S	+0.1250	-1.6250	-0.8501
Strong sign	two tail	5	3	S	-0.2155	-1.5768	-0.8019
Strong sign	center	5	3	C	-0.1875	-1.9375	-1.1626
Strong sign	two tail	5	3	C	-0.4321	-1.7471	-0.9722
Nuisance	center	2	1	S	-0.1875	+0.0938	+0.3438
Nuisance	two tail	2	1	S	+0.0156	+0.2656	+0.5156
Nuisance	center	2	1	C	-0.1719	+0.0859	+0.3984
Nuisance	two tail	2	1	C	-0.0078	+0.0391	+0.3516
Nuisance	center	3	1	S	+0.0938	+0.1641	+0.6641
Nuisance	two tail	3	1	S	+0.0000	+0.1523	+0.6523
Nuisance	center	3	1	C	+0.0859	+0.1504	+0.6973
Nuisance	two tail	3	1	C	+0.0156	+0.0430	+0.5898
Nuisance	center	5	3	S	+0.0179	+0.0718	+1.0654
Nuisance	two tail	5	3	S	-0.0461	+0.0854	+1.0790
Nuisance	center	5	3	C	+0.0108	+0.0615	+1.0612
Nuisance	two tail	5	3	C	+0.0041	+0.0298	+1.0296
Unequal tails	center	2	1	S	-0.0312	-0.8125	-1.8438
Unequal tails	two tail	2	1	S	-0.2969	-1.0000	-2.0312
Unequal tails	center	2	1	C	+0.0781	-0.7812	-2.0000
Unequal tails	two tail	2	1	C	-0.1562	-0.7031	-1.9219
Unequal tails	center	3	1	S	-0.1250	-1.1172	-1.5625
Unequal tails	two tail	3	1	S	-0.0430	-1.0820	-1.5273
Unequal tails	center	3	1	C	-0.1172	-0.9180	-1.5664

Case	Family	N	m	Rule	Δ single	Δ keep	Δ QR
Unequal tails	two tail	3	1	C	+0.0000	-0.8398	-1.4883
Unequal tails	center	5	3	S	-0.2062	-0.9900	-0.5155
Unequal tails	two tail	5	3	S	-0.2728	-1.1190	-0.6445
Unequal tails	center	5	3	C	-0.2258	-0.9056	-0.5812
Unequal tails	two tail	5	3	C	-0.2371	-0.9666	-0.6422
Radius, $p = .1$	radius	2	1	S	+0.0000	-0.0648	+3.6200
Radius, $p = .1$	radius	2	1	C	+0.0000	-0.0468	+3.6380
Radius, $p = .1$	radius	3	1	S	+0.0000	-0.0972	+3.6524
Radius, $p = .1$	radius	3	1	C	+0.0000	-0.0702	+3.6794
Radius, $p = .1$	radius	5	3	S	-0.0483	-0.1002	+3.6819
Radius, $p = .1$	radius	5	3	C	-0.0285	-0.0592	+3.7229
Radius, $p = .5$	radius	2	1	S	+0.0000	-0.5000	+2.5000
Radius, $p = .5$	radius	2	1	C	+0.0000	-0.2500	+2.7500
Radius, $p = .5$	radius	3	1	S	+0.0000	-0.7500	+2.7500
Radius, $p = .5$	radius	3	1	C	+0.0000	-0.3750	+3.1250
Radius, $p = .5$	radius	5	3	S	-0.1875	-0.9375	+2.9375
Radius, $p = .5$	radius	5	3	C	-0.0156	-0.3906	+3.4844
Radius, $p = .9$	radius	2	1	S	+0.0000	-0.0648	+2.0200
Radius, $p = .9$	radius	2	1	C	+0.0000	+0.0972	+2.1820
Radius, $p = .9$	radius	3	1	S	+0.0000	-0.0972	+2.0524
Radius, $p = .9$	radius	3	1	C	+0.0000	+0.1458	+2.2954
Radius, $p = .9$	radius	5	3	S	-0.0483	-0.1002	+2.0819
Radius, $p = .9$	radius	5	3	C	-0.0858	+0.0534	+2.2355

G.4 A DIFFERENT FAMILY AT 90 AND 95 PERCENT TARGET COVERAGE

This two-action family is distinct from the preceding nine-action example: the latter has $m = 1, N = 3$ and does not satisfy the restriction $N \le 2m$ below. It provides another exact selection-risk calculation at commonly used coverage targets, not a direct parameter extension of that counterexample.

Let the four states be as above, with $\Pr(B) = p$, $\Pr(A) = r(1-p)$, $0 \le p < 1$, $0 \le r \le 1$. States C and D divide the remaining mass. Compare fixed keep $[-1, 1]$ with update $[-1, 1 + X]$. Let $\alpha = 1/(m+1)$, $m + 1 \le N \le 2m$, and $q = N - m$. Use nonnegative joint calibration, keep on objective ties, and independent fresh final calibration of size m followed by an independent query. Write $B_s(r) = \Pr\{\mathrm{Bin}(s, r) > s/2\}$.

The key observation is that no miss in a finite history does not imply no miss in the population.

Proposition 4 (Selection event and ordinary-control equivalence)**.** *Single-pool selection chooses update with probability $(1-p)^N B_q(r)$. All-subset and raw-all selection choose it with probability $(1-p)^N B_N(r)$. The latter two selectors coincide for every mature sample and, when given the same final calibration sample, emit identical intervals on every query. If*

$$\begin{aligned} L_0 &= (3 - 2r)(1-p) + (1 + 2/\alpha)p, \\ L_1 &= (5 - 2r)(1-p) + 3p, \\ \Delta &= L_1 - 2 + (1-p)^m (L_0 - L_1), \end{aligned}$$

their common selected risk minus the single-pool risk is $\Delta(1-p)^N[B_N(r) - B_q(r)]$.

Proof. The calibration rank is the maximum. Only B has positive residual, equal to one. A calibration B gives every query score at least three. If no calibration B is present, a query B cannot be offset by the other queries because $q \le m < 1/\alpha$. Otherwise the mean query score is $1 + 2\#\{C, D\}/q$, which beats keep exactly when A is a strict majority.

For pooled counts (a, b, c) of A, B and C/D, respectively, and $b \ge 1$, put $v = \binom{N-b}{m} / \binom{N-1}{m}$, with impossible choices zero. Query-excluded subset averaging gives

$$3 + 2c/N + (2v/N)(m + b/\alpha - N) \ge 3.$$

To see this, the probability of a zero correction for a non-B query is $u = \binom{N-1-b}{m} / \binom{N-1}{m}$; use $(N-b)u = (N-b-m)v$ in the weighted four-state losses. For $b = 0$ the objective is $1 + 2c/N$. The

raw-all objective is $1 + 2b/(\alpha N) + 2c/N$. If $b \geq 1$, it is strictly above two since $N \leq 2m < 2/\alpha$; otherwise it equals the averaged objective. Thus both choose update exactly when $b = 0, a > N/2$. Their objective values may differ when $b > 0$, but their decisions do not.

Conditioning on no B yields the two binomial tails. The final correction is zero with probability $(1-p)^m$ and one otherwise. The corresponding population losses are L_0, L_1, with $L_0 - L_1 = -2 + 2p/\alpha$. Multiplying the action risk difference by its selection probabilities proves the last statement. □

For $r = 3/4$, $p = 0$ gives $\Delta = -1/2$, whereas $p = \alpha$ gives $\Delta = 3/2 - \alpha/2 > 0$. Thus more frequent selection can help or hurt. Strict-majority probabilities also depend on parity; we do not assert that $B_N - B_q$ has the same sign for all N, q, r. In this family, subset averaging is not a distinct calibration-aware decision algorithm.

G.5 What finite history cannot distinguish

The limitation is not confined to either selector. A rare state can leave the observed history unchanged while reversing the preferable action.

Proposition 5 (Two-law minimax regret)**.** *Let P_0 have $p = 0, r = 3/4$ and let $P_1 = (1-\alpha)P_0 + \alpha\delta_B$. For any possibly randomized selector of the preceding two fixed actions, based only on N mature IID observations, define regret relative to the better fixed action under each law, averaging fresh final calibration and query loss. Put $a = 1/2$, $d = 3/2 - \alpha/2$ and $w = (1-\alpha)^N$. Then*

$$\inf_{\pi} \max_{i \in \{0,1\}} \mathrm{Regret}_{P_i}(\pi) = \frac{adw}{a + dw}.$$

Proof. Update improves on keep by a under P_0 and worsens it by d under P_1. Let E be the event that the mature sample contains no B. Conditional on E, the law under P_1 is exactly P_0^N, and $P_1(E) = w$. If u is the probability of updating under P_0, the probability under P_1 is at least wu. The two regrets are therefore $a(1-u)$ and at least dwu. Minimizing their maximum over $0 \leq u \leq 1$ gives the bound. Equality is attained by updating with probability $a/(a + dw)$ on E and keeping otherwise. □

This is a classical two-law indistinguishability argument with explicitly computed finite-calibration risks. The equality describes a known two-law experiment, not an implementable universally optimal update. Regression that emits other endpoints is outside this action class. As α decreases with $N \leq 2m = 2/\alpha - 2$, the rare state's loss grows as $1/\alpha$; the obstruction is not a uniform-in-α bounded-loss lower bound.

Observable quantities and information limits. α, m, N are design quantities. State labels are available only after outcomes mature; p, r are unknown population parameters. Estimating zero miss mass on a history with no B is not a safe estimate of zero future risk. Standard binomial confidence regions can express uncertainty under the four-state IID model, but do not remove misspecification or dependence. Neither proposition supplies a general learned update rule.

Complete common-target comparison. The frozen matrix contains 24 artificial populations: two error levels $.10, .05$, their minimum finite calibration sizes $9, 19$, two mature sample sizes $m + 1, 2m$, and six specified probability laws. Seven controls yield 168 method configurations. The all-subset/raw-all decision lowers risk relative to single-pool selection in eight configurations and raises it in sixteen. These are exhaustive specified cases, not sampled domain wins. The known shifts zero and three are translation-equivariance metadata, not extra experiments or fitted nuisance parameters. All reported method expectations were evaluated as exact rational sums; the frozen plan had allowed float64 for the regression control. This arithmetic strengthening changes neither the statistical object nor the selected family.

The conservative Hoeffding reference selects keep in every configuration: its margin exceeds the largest possible empirical gain. Its safety here is therefore trivial. A query bound conditional on an internal calibration pool is not a bound for R_m after independently redrawing that pool. Joint QR uses empirical conditional quantiles on the two feature values and a pooled fallback when a feature

value is absent. It has the same N labels but a larger endpoint class. Full score, coverage, width, maximum support loss and action probabilities accompany Table 33.

Table 33: Exact finite-population comparison. All-subset and raw-all selectors are pathwise identical in this family; their common score difference from single-pool selection is shown. Negative is favorable. p is miss-state mass and r is the conditional probability of state A given no miss. Translations preserve all losses and do not constitute repetitions.

α	N, m	p	r	Δ	Avg. − single	QR score
0.10	10,9	0.2500	0.333	+2.2253	-0.03218	2.7448
0.10	10,9	0.0000	0.750	-0.5000	-0.08594	2.2857
0.10	10,9	0.0250	0.750	+0.2931	+0.03911	3.3193
0.10	10,9	0.0500	0.750	+0.8448	+0.08693	3.7895
0.10	10,9	0.1000	0.750	+1.4500	+0.08690	4.0043
0.10	10,9	0.2000	0.750	+1.6684	+0.03079	3.8621
0.10	18,9	0.2500	0.333	+2.2253	-0.00127	2.0777
0.10	18,9	0.0000	0.750	-0.5000	-0.01479	1.1657
0.10	18,9	0.0250	0.750	+0.2931	+0.00550	2.5968
0.10	18,9	0.0500	0.750	+0.8448	+0.00993	3.0364
0.10	18,9	0.1000	0.750	+1.4500	+0.00644	3.0677
0.10	18,9	0.2000	0.750	+1.6684	+0.00089	2.9910
0.05	20,19	0.2500	0.333	+2.0338	-0.00191	2.0389
0.05	20,19	0.0000	0.750	-0.5000	-0.11807	1.0375
0.05	20,19	0.0125	0.750	+0.3126	+0.05740	2.4788
0.05	20,19	0.0250	0.750	+0.8694	+0.12372	3.0922
0.05	20,19	0.0500	0.750	+1.4750	+0.12486	3.2940
0.05	20,19	0.1000	0.750	+1.7202	+0.04938	2.9823
0.05	38,19	0.2500	0.333	+2.0338	-1.96×10^{-6}	2.0002
0.05	38,19	0.0000	0.750	-0.5000	-0.00406	0.5497
0.05	38,19	0.0125	0.750	+0.3126	+0.00157	2.1217
0.05	38,19	0.0250	0.750	+0.8694	+0.00270	2.5215
0.05	38,19	0.0500	0.750	+1.4750	+0.00171	2.4106
0.05	38,19	0.1000	0.750	+1.7202	+0.00025	2.1402

G.6 When choosing on the same calibration data helps

Let a finite candidate set $\mathcal{A}$ be fixed in advance, let R_m be the fresh-calibration interval risk of equation 1 (Section 2), and let $\widehat{a}(D)$ minimize any estimate $\widehat{R}$ computed from the mature sample D, ties broken by a stated rule. Fix a baseline a_0 and a reference $a^\star$, and write $\Delta_a = R_m(a) - R_m(a^\star)$, $g^\star = R_m(a_0) - R_m(a^\star)$ and $\pi_a = \Pr\{\widehat{a} = a\}$. The reversal indicator of a candidate against the reference is $\rho_a(D) = \mathbf{1}\{\widehat{R}(a) < \widehat{R}(a^\star)\} + \theta_a \mathbf{1}\{\widehat{R}(a) = \widehat{R}(a^\star)\}$, where $\theta_a = 1$ if a priority order ranks a above $a^\star$ and 0 if it ranks it below, $\theta_a = 1/2$ under uniform ties, and $\theta_a = 1$ for an arbitrary rule; $\beta_a = \mathbb{E}\rho_a(D)$. Let $V(a^\star) = \max\{\sum_a (\Delta_a)_+ x_a : 0 \le x_a \le \beta_a,\ \sum_a x_a \le 1\}$, which is at most $\sum_a (\Delta_a)_+ \beta_a$ and equals it when the worse candidates' β_a sum to at most one.

Lemma. $\Pr\{\widehat{a} = a \mid D\} \le \rho_a(D)$ for every $a \ne a^\star$, hence $\pi_a \le \beta_a$. If $\widehat{R}(a) > \widehat{R}(a^\star)$ the candidate is not a minimizer; if it is smaller the right side is one; if the two are tied, both or neither are minimizers, and a priority order can pick a only when it ranks a first, uniform ties pick it with probability at most $1/2$, and any rule with probability at most one.

Theorem 4 in full. (i) $R_m(a_0) - \mathbb{E}R_m(\widehat{a}) = g^\star - \sum_{a \ne a^\star} \Delta_a \pi_a$. (ii) $\mathbb{E}R_m(\widehat{a}) \le R_m(a^\star) + V(a^\star)$. (iii) If $V(a^\star) < g^\star$, selection lowers the selected risk below $R_m(a_0)$ by at least $g^\star - V(a^\star)$. (iv) If $\Pr\{\widehat{a} = a \mid D\} = \rho_a(D)$ for every $a \ne a^\star$ and $a^\star$ is a best candidate, selection improves on a_0 if and only if $V(a^\star) < g^\star$, and the improvement equals $g^\star - V(a^\star)$. (v) For any risks and any values of the worse candidates' β_a, a tie-free law of estimates attains $R_m(a^\star) + V(a^\star)$, so no bound that uses only this information is smaller. Item (i) is the identity $\mathbb{E}R_m(\widehat{a}) = \sum_a \pi_a R_m(a)$; item (ii) follows because $(\pi_a)_{a \ne a^\star}$ is feasible for the program defining $V(a^\star)$ by the lemma; item (iv) takes expectations of the equality; item (v) assigns consecutive intervals of the unit interval to the worse

candidates, in nonincreasing order of their gaps, with lengths given by the greedy solution of the program, so that each is the unique minimizer on its interval; where the unit interval runs out before a candidate's β_a is used, the candidate is placed strictly between the minimizer and $a^\star$ on further points of the other candidates' intervals, with all values distinct, until its reversal probability reaches β_a; any unassigned mass makes $a^\star$ the minimizer.

A family in which the bound is an identity. Take signed joint calibration, $\alpha = 1/5$, $m = 4$ and the support $A = (x{=}-1, y{=}-1)$, $C = (+1, -1)$, $D = (+1, +1)$, and write $[a, b]; [c, d]$ for the interval $[a, b]$ at $x = -1$ and $[c, d]$ at $x = +1$. The unchanged interval $[-1, 1]; [-1, 1]$, a moderate candidate $[-2, 0]; [-2, 1]$ and an aggressive candidate $[-2, 0]; [-2, 0]$ satisfy, for every calibration pool and query, a loss for the moderate candidate equal to the average of the other two minus one on A. With the moderate candidate as reference and ties broken in the priority order unchanged, moderate, aggressive, every selection coincides with pairwise reversal. For every law, every number of queries and every all-subset sample size the selected risk therefore equals $R_m(a^\star) + \sum_{a \neq a^\star} \Delta_a \beta_a$, and item (iv) applies whenever the moderate candidate is best. The order matters: ranking the aggressive candidate above the moderate one breaks the coincidence at three-way ties (for example with ten queries) and, with one query, turns the first law below from an improvement into a loss of $203069511/500000000$. With a single pool of four and one query, the law $(p_A, p_C, p_D) = (13/20, 1/4, 1/10)$ gives $g^\star = 3061/10000$ and an improvement of exactly $35449083/2000000000$, while $(3/5, 3/10, 1/10)$ gives $g^\star = 2561/10000$ and a loss; the sign changes near $p_C = 0.2666$, so no condition of the form $cV < g^\star$ with $c < 1$ is valid.

The construction of Theorem 3. There the unchanged interval is the unique best candidate, with risk 2 against $13/4$ and 4, so $g^\star \leq 0$ for every reference and the condition cannot hold. With the unchanged interval as reference and the declared rule, selection coincides with pairwise reversal, and the excess risk equals $\sum_a \Delta_a \beta_a$: $4 \cdot \frac{5}{4} \cdot \frac{3}{64} = \frac{15}{64}$ for the single pool and $4 \cdot \frac{5}{4} \cdot \frac{7}{64} = \frac{35}{64}$ for averaging, reproducing $143/64$ and $163/64$. For each one-tail action the single pool reverses strictly with probability $3/64$ and ties with probability $12/64$; averaging reverses strictly with probability $7/64$ and never ties. Ties cost nothing under the declared rule, because the unchanged interval is first and best, so converting tie mass into strict reversals is the whole increase. The paired variances fall at the same time, which is the point of the construction.

A strict improvement, and noise that more queries cannot remove. For the nine actions of Section K, recalibrated by signed joint calibration at $\alpha = 1/5$ with pools of four, and the law $(9/10, 0, 1/20, 1/20)$ on $(x, y) \in \{(-1, -1), (-1, 1), (1, -1), (1, 1)\}$, the best candidate shrinks the upper endpoint where the outcome is always low, with risk $1920039/1600000$ against 2 for the unchanged interval. With seven mature observations the condition certifies an improvement for both selectors and both tie rules; under the declared rule the certified margin is about 0.548 for all-subset selection and 0.512 for the single pool. A two-state witness separates the selectors under the same calibration. Its candidates are the unchanged interval, $[0, 2]; [-2, 2]$ and $[0, 0]; [-2, 1]$, with support $(-1, +1)$ and $(+1, -1)$ at probabilities $1/12$ and $11/12$; the second is best, with risk $11/6$, and the third, with risk $18191/6912$, is conditionally best on pools without the rare state but not on average. It makes single-pool selection worse than the unchanged interval for every number of queries from 1 to 400 and in the limit, where its risk is $343570271/143327232 \approx 2.397$ against 2: the error lives in the one calibration pool, and more queries do not reach it. All-subset selection averages over pools, and exact reversal probabilities certify it better than the unchanged interval for every mature sample size from 36 to 400.

A limiting gap is not a necessary finite-sample condition. Under a unique conditional optimum for almost every internal calibration pool, single-pool selection converges to the mean fresh-calibration risk of that conditional optimum. All-subset selection converges to the minimum fresh risk. A strictly positive difference between these limits suffices for eventual strict improvement, but is not necessary. Let $Y = 0$, $X \in \{0, 1\}$ with $\Pr(X = 0) = p \in (0, 1)$, and compare $[0, 0]$ with baseline $[0, X]$, preferring the baseline on ties. Every joint residual is zero. At any valid finite rank with $N > m \geq 1$, single-pool selection has fresh risk $(1 - p)p^{N-m}$, whereas all-subset selection has risk $(1 - p)p^N$: each mature observation has equal total query weight. The latter is strictly smaller for every finite $N > m$, although both limits and every pool's optimum risk are zero. This distinction does not change Theorem 4 or its attainment construction.

H MORE QUERIES CAN LEAVE A CALIBRATION ERROR UNRESOLVED

An internal calibration pool is a random part of the procedure being assessed. More queries can estimate its conditional risk accurately without estimating the risk of a fresh calibration. This section makes that distinction explicit with continuous outcomes, then tests ordinary allocation controls and two pilot heuristics. The construction and the comparisons use different candidate families and answer different questions.

H.1 A STRICT-RANKING CONSTRUCTION AT FINITE CALIBRATION SIZE

Let $m \geq 1$, $\alpha = 1/(m+1)$ and $p = 1/[2(m+2)]$. A rare covariate state has probability p and outcome 1; the other state has outcome -1. Fix the three raw interval functions below before seeing the selection observations.

Forecast	Rare state	Common state
Unchanged	$[-1, 1]$	$[-1, 1]$
Moderate	$[0, 2]$	$[-2, 2]$
Aggressive	$[0, 0]$	$[-2, 1]$

Use unit scales, signed joint calibration at rank m, and midpoint collapse. Write $R_m(a)$ for risk with a fresh independent size-m calibration pool and a fresh query. Add independent observation noise with conditional distribution supported on $[-\epsilon, \epsilon]$, where $\epsilon \geq 0$. Positive-width uniform noise makes outcomes continuous. Candidate functions remain fixed.

Proposition 6 (A finite-query error floor)**.** *Suppose $\epsilon < 1/4$, $2(8m+10)\epsilon < p$ and $\epsilon < 1/[2(m+1)]$. Select the forecast with lowest average interval score on $q \geq 1$ IID queries using one independent internal pool of size m. Use any tie rule and a fresh independent final calibration pool. Set*

$$D_m = (8m+10)\epsilon, \qquad \Delta_m = 3 - p - 3(1-p)^m, \qquad k_q = \left\lceil \frac{q}{1 + 2(m+1)(1+\epsilon)} \right\rceil - 1.$$

Then the moderate forecast uniquely minimizes fresh risk, and

$$\mathbb{E}\{R_m(\widehat{a}) - R_m(a_{\text{moderate}})\} \;\geq\; (1-p)^m(\Delta_m - 2D_m)\Pr\{\text{Bin}(q,p) \leq k_q\} > 0. \tag{17}$$

For each fixed m, the binomial probability tends to one as q increases. The choice $\epsilon = p/[4(8m+10)]$ satisfies all conditions.

Proof. First take zero noise. Without a rare calibration observation, the conditional query risks, in the order above, are $(2, 2-2p, 2-3p)$. With at least one rare observation they are $(2, 2-2p, 5-3p)$. The smallest winning margin is p. The moderate forecast's fresh risk is $2-2p$, whereas aggressive has excess $\Delta_m \geq 2p$, since $(1-p)^m \leq 1-p$. Unchanged has excess $2p$.

Couple the noisy and noiseless observations. Each residual and its order statistic change by at most ϵ. Midpoint collapse is a positive-part operation on the half-width, so each issued endpoint changes by at most ϵ. Width changes by at most 2ϵ and each outside distance by at most 2ϵ, including the query noise. The loss changes by at most $(2 + 8/\alpha)\epsilon = D_m$. Each paired conditional or fresh-risk difference changes by at most $2D_m$, preserving the strict winners.

On an all-common internal pool, let t be its smallest noise value. The three calibration radii are $-t, -1-t, -1-t$. For a common query, aggressive has exactly one unit less loss than either competitor: lower endpoints and lower-tail penalties coincide, and the smaller upper endpoint is still above the outcome. For a rare query it collapses to $[0, 0]$, with loss at most $2(m+1)(1+\epsilon)$; competitor losses are nonnegative. If K of the q queries are rare, aggressive-minus-competitor total loss is therefore at most $-q + \{1 + 2(m+1)(1+\epsilon)\}K$. It is strictly negative for $K \leq k_q$. The internal-pool event has probability $(1-p)^m$ and is independent of $K \sim \text{Bin}(q, p)$. Aggressive has fresh excess at least $\Delta_m - 2D_m > 0$; other selections have nonnegative excess. This proves the bound, and $K = 0$ proves its positivity for every finite q. Finally, $1/[1 + 2(m+1)(1+\epsilon)] > p$, so the binomial law of large numbers gives the limit. All queries share the internal pool; the final pool is independent of selection. □

This is a family indexed by m, not one fixed population for all m. In particular $m = 9$ and 19 give nominal coverage 90 and 95 percent. The allowed noise radius decreases with m. The result does not assert that each extra query increases regret, that more pools always help, or that the same failure survives adding a well-fitted candidate. It removes dependence on a winning tie and on discrete outcomes from this fixed-pool example, not the general limitations of model selection.

H.2 THE ORDINARY CONTROLS CAN REMOVE THE DIFFICULTY

We evaluated ten fixed synthetic settings, each with 128 independent fitting, selection and final-evaluation repetitions. The settings vary outcome noise, scale, nominal coverage, rare-state probability, signed versus nonnegative calibration and paired residual ranks. Every candidate bank additionally includes a state-stratified empirical quantile forecast fitted on 128 independent observations. This fourth candidate is absent from the proposition.

Every rule receives an available budget of $12m + 48$ unique selection labels. Fixed allocations use one, two, four or eight disjoint pools, with remaining labels as queries. A complete query-excluded subset average and an average of 32 random splits can use the entire budget. The two pilot rules pay $4m + 16$ labels to choose an allocation, leaving $8m + 32$ for selection. One minimizes estimated paired variance; the other minimizes regret estimated by resampling the pilot. Both are ordinary development heuristics, without a risk guarantee. All selected forecasts receive a fresh independent size-m final pool. The unchanged and fixed-quantile controls use zero selection labels to decide their action; the quantile forecast still requires its 128 fitting labels. Shared raw-score diagnostics are not decision costs. The accompanying cost correction separates these quantities and leaves unmeasured independent runtimes unspecified; the original harness timers do not establish efficiency rankings.

Support of the conservative control. In all ten settings, $|Y| \le 5/4$, raw endpoints lie in $[-2, 2]$, scales lie in $[1, 3]$, and $\alpha \ge 1/20$. The fitted empirical quantiles inherit the outcome support, including the empty-state fallback. Joint residuals and every finite calibration radius have absolute value at most $13/4$. Nonnegative truncation preserves this bound; issued endpoints before and after midpoint collapse lie in $[-47/4, 47/4]$. Ordered endpoints allow at most one positive outside penalty, whose distance is at most 13. Therefore every issued loss is at most $47/2 + 26/\alpha \le 543.5$. The original conservative rule retains its valid, looser bound of 4000; no threshold is changed after observing results. Overflow ranks have no finite bound and are excluded by these settings.

Table 34: Adding a fitted quantile control removes most of the constructed selection difficulty. Entries are expected fresh-pool interval scores averaged over 128 independent repetitions. All ten settings are retained.

Setting	Fitted QR	Repeated	Complete	Variance	Bootstrap
Discrete outcomes	0.0000	0.0000	0.0000	0.0000	0.0000
Small continuous noise	0.0005	0.0005	0.0005	0.0005	0.0005
Wide continuous noise	0.4950	0.4950	0.4950	0.4950	0.4950
Variable scale	0.4967	0.4967	0.4967	0.4967	0.4967
Nominal 90 percent	0.0005	0.0005	0.0005	0.0005	0.0005
Nominal 95 percent	0.0005	0.0005	0.0005	0.0005	0.0005
Rare miss	0.0636	0.0636	0.0636	0.0636	0.0937
Nonnegative correction	0.4714	0.4714	0.4714	0.4714	0.4714
Aligned ranks	0.4924	0.4843	0.4845	0.4851	0.4853
Opposed ranks	0.4923	0.4843	0.4838	0.4859	0.4856

The full twelve-rule results include unchanged, fixed quantile, raw-loss and conservative controls. None of the 30 prespecified paired comparisons of the bootstrap pilot against variance allocation, repeated splits and one fixed pool survives Holm correction. This is not an equivalence result. The table does not test the three-candidate theorem after silently changing its candidate family.

H.3 A FIXED-BUDGET HISTORICAL COMPARISON

We next reused the saved fit-only linear quantile, clipped-feature quantile and Student predictive forecasts from the energy and traffic histories. No model was refitted. Every selector receives the

same three forecasts and an available budget of 78 mature historical selection labels, with nine final-calibration labels at nominal 90 percent coverage. The pilot pays 26 labels and chooses one, two or four pools from the remaining 52. Other rules may use all 78 labels. The final pool and test outcomes never select an allocation or forecast. This is a new small-pool comparison, not a rerun of the larger-pool online comparison in Appendix F.2. Fixed forecasts and the unconditional conservative fallback need no selection labels to choose an action. Their historical fitting costs are not zero; fitting time was not remeasured. No population loss bound is assumed here.

Table 35: The pilot rules do not consistently improve on ordinary controls. Final interval score is averaged equally over days within a seven-day episode, then equally over episodes. Energy uses Wh and traffic uses vehicles/hour. Zero and one denote hypothetical extra native ticks before labels become available. These are four fixed descriptive panels, not four new populations.

Rule	Energy 0	Energy 1	Traffic 0	Traffic 1
Linear quantile forecast	300.3930	323.6068	2511.6836	4171.8238
Clipped-feature quantile forecast	298.2566	323.4982	2511.6836	4179.3300
Raw-loss selection	300.2692	323.8966	2912.6869	4176.6947
One fixed pool	300.3925	323.8781	2963.9263	4521.2492
Two fixed pools	300.4404	323.8781	2912.6869	4176.6947
Four fixed pools	300.3925	323.9371	2912.6869	4724.9408
Complete subset average	300.2470	323.8966	2511.6836	4176.6947
Repeated split average	300.2470	323.8781	2511.6836	4176.6947
Variance-selected allocation	308.0543	339.9126	2912.6869	4169.6866
Bootstrap-selected allocation	307.9764	323.8622	2912.6869	4170.1465
Conservative fallback	300.3930	323.6068	2511.6836	4171.8238

All 40 episode–delay requests are retained: 32 qualify, six inherit source ineligibility, and two have fewer than 78 selection labels. There are ten eligible episodes per energy setting and six per traffic setting. No replacement window is introduced. The archives were already opened development data; actual publication times are unknown, and dependent chronological samples do not satisfy the proposition's IID assumptions. The complete companion reports episode-level paired differences, block lengths one, two and three, all raw and issued scores, coverage, width, outside penalties, choices and operation counts. Its small-sample block ranges are exploratory, not proof of general superiority.

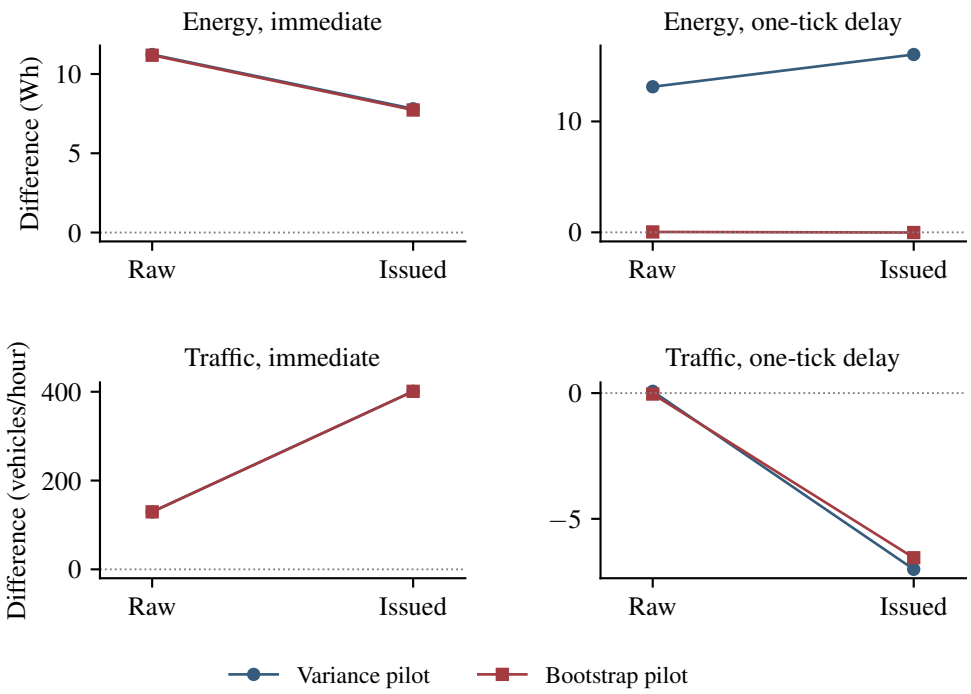

Figure 3: Pilot allocation has no consistent raw-to-issued advantage over ordinary repeated-split selection. Positive differences favor repeated splits. Each line compares the same selected forecast before and after its final calibration, using the source-specific equal-day, equal-episode means of Table 35. These are point contrasts, not confidence intervals or independent-source replications.

Table 36: Coverage and score components for fixed forecasts and principal selection controls. Choices count linear, Student and clipped-feature forecasts, respectively. A lower mean score need not imply better coverage.

Panel	Rule	Coverage	Width	Outside	Choices
Energy 0	Linear QR	0.8631	200.8252	99.5677	[10, 0, 0]
Energy 0	Clipped QR	0.8619	197.7957	100.4609	[0, 0, 10]
Energy 0	Repeated split	0.8630	200.4935	99.7535	[6, 0, 4]
Energy 0	Complete subsets	0.8630	200.4935	99.7535	[6, 0, 4]
Energy 0	Variance pilot	0.8700	197.5162	110.5381	[6, 1, 3]
Energy 0	Bootstrap pilot	0.8703	197.5414	110.4350	[7, 1, 2]
Energy 1	Linear QR	0.8747	193.5379	130.0689	[10, 0, 0]
Energy 1	Clipped QR	0.8748	192.5892	130.9090	[0, 0, 10]
Energy 1	Repeated split	0.8748	193.1048	130.7734	[4, 0, 6]
Energy 1	Complete subsets	0.8747	193.0971	130.7995	[5, 0, 5]
Energy 1	Variance pilot	0.8882	212.3683	127.5443	[3, 1, 6]
Energy 1	Bootstrap pilot	0.8751	193.2555	130.6067	[6, 0, 4]
Traffic 0	Linear QR	0.8927	2024.4231	487.2605	[6, 0, 0]
Traffic 0	Clipped QR	0.8927	2024.4231	487.2605	[0, 0, 6]
Traffic 0	Repeated split	0.8927	2024.4231	487.2605	[6, 0, 0]
Traffic 0	Complete subsets	0.8927	2024.4231	487.2605	[6, 0, 0]
Traffic 0	Variance pilot	0.8927	2434.5473	478.1395	[5, 1, 0]
Traffic 0	Bootstrap pilot	0.8927	2434.5473	478.1395	[5, 1, 0]
Traffic 1	Linear QR	0.8744	3454.7892	717.0346	[6, 0, 0]
Traffic 1	Clipped QR	0.8724	3465.0868	714.2431	[0, 0, 6]
Traffic 1	Repeated split	0.8744	3464.4802	712.2145	[4, 0, 2]
Traffic 1	Complete subsets	0.8744	3464.4802	712.2145	[4, 0, 2]
Traffic 1	Variance pilot	0.8744	3457.5565	712.1301	[4, 0, 2]
Traffic 1	Bootstrap pilot	0.8744	3457.9320	712.2145	[5, 0, 1]

I EARLIER ENDPOINT AND SELECTION ANALYSES

The following analyses retain their original assumptions and comparisons. They provide context for the fresh-calibration question; they are not evidence that the new risk-comparison rule improves every forecasting task.

I.1 THREE PREDICTION OBJECTS

The raw forecast, selection interval and issued interval need not improve together (Figure 4). Let H_t be the history available at issue time, and let x be a feature of that history. An upstream model supplies ordered raw endpoints $L^0(x), U^0(x)$ and a fixed positive scale $c(x)$. An action a changes the raw endpoints. An internal calibration sample C_{in} defines the interval used to choose a on separate selection rows. Only afterward does a final sample C_{fin} calibrate the chosen action. We score that last interval, not the selection interval.

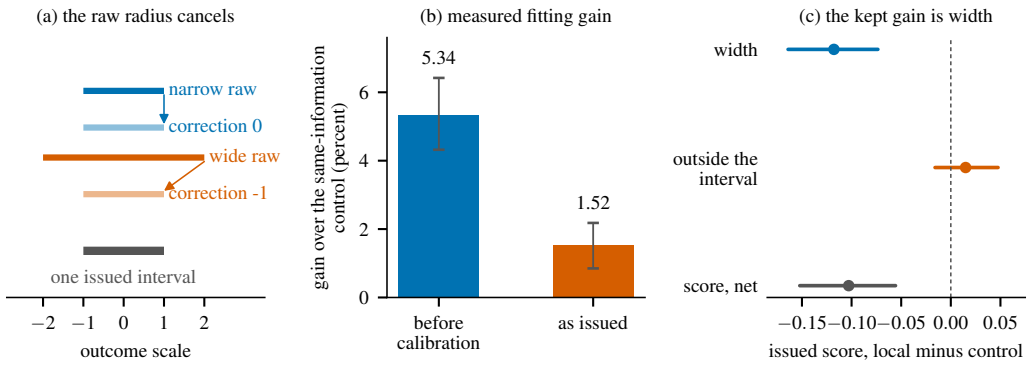

Figure 4: Earlier analyses of raw and issued intervals. (a) Signed joint calibration removes the raw radius: with unit scales, $\alpha = 1/2$ and one calibration point at 1, radii of 1 and 2 take corrections 0 and -1 and are issued as the same interval. This panel is exact arithmetic. (b) A locally fitted predictor beats a parameter-identical control on the same information by 5.34 percent before calibration and 1.52 percent as issued, over three seeds and nine domains. (c) The issued contrast splits with no cross term: the width part carries all of it, and the outside-the-interval part runs the other way with an interval containing zero. Bars are 95 percent series-within-domain bootstrap intervals.

For signed joint calibration on a pool C of size n, define

$$r_i = \max\{(L^0(x_i) - y_i)/c_i, (y_i - U^0(x_i))/c_i\},$$
$$k = \lceil (n+1)(1-\alpha) \rceil, \qquad q = r_{(k)}, \qquad I_C(x) = [L^0(x) - c_x q, U^0(x) + c_x q]. \tag{18}$$

If $k > n$, the interval is unbounded; replacing this rank by the largest observed score is a different procedure. When signed shrinking reverses finite endpoints, our specified emission rule collapses them to their midpoint. Nonnegative calibration instead truncates q at zero. Separate-tail calibration uses two ranks at miscoverage allocation $\alpha/2$. Each operation is stated where used.

For ordered finite endpoints, the interval score is

$$S_\alpha([L, U], y) = U - L + \frac{2}{\alpha}(L - y)_+ + \frac{2}{\alpha}(y - U)_+.$$

For an action fixed before final calibration, its fresh-calibration risk is $R_m(a) = \mathbb{E}_{C_m, Z} S_\alpha(I_{a, C_m}(X), Y)$. A selector based on mature data D is evaluated by $\mathbb{E}_D R_m(\hat{a}(D))$. A smaller internal objective alone says neither that this expectation decreases nor that calibration coverage holds under temporal dependence.

J What Calibration Cancels

Some raw improvements are invisible in the issued endpoints. The following statements hold with the pool, labels, positive scales and finite rank fixed; they are properties of the calibration map, not estimation-rate guarantees.

Proposition 7 (Radius cancellation)**.** *Replace every raw calibration and query interval by* $[L_i - c_i t, U_i + c_i t]$ *for a common scalar* t *that preserves raw ordering. Under signed joint calibration, every issued interval is unchanged, including after midpoint collapse.*

Every residual and its order statistic decrease by t, cancelling the raw radius change. This is close to a familiar property of split conformal prediction: a change that preserves the order of the calibration scores selects the same calibration row (Lei et al., 2018). The same row does not by itself fix the endpoints, which use its residual value; here the raw-endpoint change and the selected-residual change are equal and opposite, so the issued interval is identical, through score ties. The new formal content of this section is Proposition 8, which says which directions of a fitted adjustment survive and how many there are. Nonnegative truncation generally breaks the identity. Signed separate-tail calibration similarly cancels two common standardized endpoint offsets. The cancellation is exact on the data as well as on paper: adding a common radius shift of 0.25 to a fitted variant's raw endpoints on all 153 development series leaves every issued endpoint unchanged to 0.0. A reader can therefore improve a raw interval in this direction and issue exactly the interval they started with.

Endpoint geometry alone does not price the outcomes missed by an interval.

Proposition 8 (Local endpoint equivalence)**.** *Inside a region of strict raw ordering, residual-tail choice and calibration rank, let raw endpoints be affine in* θ*, and assume issued endpoints do not collapse. Stack their derivatives on an observation set* B *in* A_B*. Let* $g_B = (-c_B^\top, c_B^\top)^\top$ *and let* $w^\top$ *be the derivative of the selected calibration residual. Then*

$$J_B = A_B + g_B w^\top.$$

Within this affine region, a displacement leaves the issued endpoints on B *unchanged exactly when it belongs to* $\ker J_B$*. If* B *contains the calibration rows, let* $h^\top$ *select the active residual derivative from this full raw endpoint vector. Then* $h^\top g_B = -1$*,* $J_B = PA_B$ *with* $P = I + g_B h^\top$*, and*

$$\dim\ker(PA_B) = \dim\ker A_B + \dim\big(\operatorname{im} A_B \cap \operatorname{span} g_B\big).$$

Thus a model direction can disappear because it already leaves raw endpoints unchanged, or because it moves only their common standardized radius. The dimension formula applies to the stated observation set containing the calibration rows, not automatically to test-only endpoints. For nonlinear heads the Jacobian describes first-order changes only. Ties, rank boundaries and collapse require their own regions. Appendix C gives the derivation.

Cancellation does not justify deleting a nuisance parameter from fitting. For observations $(1, 4), (2, 5)$, regression with an intercept gives slope 1; forcing zero intercept gives $14/5$. Calibration can remove a level offset without repairing this slope error. Joint ordinary regression already handles this example. Nor does endpoint equality imply equal path distributions or energy scores.

K When Choosing Hurts and When It Helps

More accurate assessment of one calibration pool need not identify the best forecast for a fresh pool. A three-forecast construction has conditional scores $(2, 11/6, 7/4)$ when a four-observation pool misses a rare state, but fresh-pool scores $(2, 11/6, 18191/6912)$ at nominal 80 percent coverage: the apparent winner becomes worst. Proposition 6 extends this failure to continuous outcomes and common coverage levels, with a positive error floor for every finite query count. A different construction below shows why even reducing paired estimation variance need not help.

Theorem 3 (Averaging can increase selected risk)**.** *Let* X, Y *be independent and uniform on* $\{-1, 1\}$*, with three IID mature observations and* $\alpha = 1/2$*. The nine raw actions sort the endpoints* $[-1 + aX, 1 + bX]$*,* $(a, b) \in \{-1, 0, 1\}^2$*, and use nonnegative joint calibration. Compare selection using the first observation to calibrate and the other two to score, with selection using the average over all six ordered calibration/query pairs. Break objective ties by* $a^2 + b^2$*, then lexicographically. With one*

Table 37: Exact selection counts among 64 equally likely mature samples. Averaging chooses harmful one-tail actions more often despite lower estimation variance.

Action	Fresh risk	Single pool	All subsets
Unchanged	2	52	36
Each of four one-tail actions	$13/4$	3	7
Each of four two-tail actions	4	0	0

fresh independent calibration observation and one fresh query, averaging increases expected selected risk from $143/64$ *to* $163/64$ *and harmful-selection probability from* $12/64$ *to* $28/64$*. The fixed-action variance decreases by* $5/24$ *for each one-tail action and by* $1/3$ *for each two-tail action; it remains zero for the unchanged interval. The same reductions hold for paired differences from that interval.*

Table 37 gives every count the construction turns on. The direction depends on that tie rule and we state both sides. The single-pool objective is tied at its minimum in 44 of the 64 mature samples, so the convention decides most selections. Under the declared rule, which prefers the smaller action and therefore the unchanged interval, averaging is worse. Under uniform-random tie-breaking the comparison reverses, to $271/96$ against $175/64$. The construction is therefore a counterexample and not a ranking: lower estimator variance does not imply lower selected risk, because there is a deterministic and natural convention under which it fails (Appendix G).

The proof enumerates the four observation states and their 64 triples (Appendix G). It is a conformal-selection counterexample to a universal implication, not a claim that the risk-estimation versus selection distinction is new.

Theorem 4 (When choosing helps)**.** *Fix candidates in advance, a baseline* a_0*, any risk estimate computed from the mature sample and any tie rule. For a reference* $a^\star$ *let* $\Delta_a = R_m(a) - R_m(a^\star)$*, and let* β_a *be the probability that the estimate of* a *is below that of* $a^\star$*, plus* θ_a *times the probability that they tie, where* θ_a *is* 1 *or* 0 *as a priority order ranks* a *above or below* $a^\star$*,* $1/2$ *for uniform ties and* 1 *for any other rule. Then* $\mathbb{E}R_m(\widehat{a}) \le R_m(a^\star) + \sum_{a:\Delta_a>0} \Delta_a\beta_a$*, so selection lowers the selected risk below* $R_m(a_0)$ *whenever* $R_m(a_0) - R_m(a^\star) > \sum_{a:\Delta_a>0} \Delta_a\beta_a$*. When the* β_a *of the worse candidates sum to at most one, no bound that uses only the gaps and the* β_a *is smaller (Appendix G.6 gives the sharp bound otherwise), and at* $\alpha = 1/5$ *a signed joint family attains it for every valid sample size and every law under which its reference candidate is best, under the stated keep–moderate–aggressive priority order.*

The bound is a union bound over pairwise reversals; what it adds here is which reversals count. Variance does not enter, and a tie counts only as the rule resolves it. With the unchanged interval as reference, Theorem 3 is an equality case: averaging raises each one-tail action's strict reversal probability from $3/64$ to $7/64$ by resolving ties that the declared rule gave to the unchanged interval, the two-tail actions never reverse strictly, and this is the whole increase in excess risk from $15/64$ to $35/64$ (Appendix G.6).

Exact search over the finite arrangement of calibration contacts is the same estimator as ordinary search over those events, and its positive and negative witnesses use different action families. At $\alpha = 1/(m+1)$ and $m+1 \le N \le 2m$ a two-action family makes all-subset and raw-loss selection identical, and two laws differing by a rare miss give every selector restricted to those actions positive minimax regret (Propositions 4 and 5). Appendix B gives a complementary sufficient condition: globally nearby endpoints with a common bounded scale admit finite-label score comparisons without a bound on the response.

L RETAINED FORECASTING PROTOCOLS

L.1 MEASURING THE PRICE

Three things are measured here, in the order a change to a forecast meets them. A locally fitted predictor beats a control that sees exactly the same information. Calibration then takes most of that gain away. What is left is narrower intervals and not better coverage. The rest of this section says how each was measured and against what.

L.2 PROTOCOL AND COMPARISON OBJECTS

We distinguish fitting, calibration, geometry and transfer rather than combine them into one experiment. All comparisons record the available information, the emitted endpoints, the target coverage and the statistical unit. Final calibration labels do not choose a deployed candidate. Temporal order is enforced in the historical studies but is not an exchangeability guarantee.

The matched neural comparison uses Electricity, Exchange, Solar and Traffic (Lai et al., 2018), ETTh1, ETTh2, ETTm1 and ETTm2 (Zhou et al., 2021), and Weather (Wu et al., 2021). Histories and futures are normalized, clipped and resampled to 128 and 16 coordinates. Three folds each hold out three domains from training. The history-local fit and amortized control share 52,986 parameters, initialization, decoder and training settings. A four-dimensional coefficient is fitted from observed transition residuals instead of directly using the amortized coefficient:

$$\widehat{\beta} = (\Phi^\top \Phi/n + \lambda I)^{-1}(\Phi^\top e/n + \lambda\beta_0). \tag{19}$$

The representation is trained through this fit; the decoder is a normalized triangular Student-t law with five degrees of freedom and variance floor .01. Ridge residualization of its basis is generally not exact orthogonal projection; its finite-ridge identity and conditional error bound are in Appendix C.3. These components do not introduce local regression as a learning principle.

The earlier Gaussian version exceeded its then-binding width-ratio limit: $1.0277 > 1.02$. A subsequent declared protocol changed the decoder and treated width ratio at most 1.05 as a reported diagnostic; it did not retroactively make the Gaussian attempt pass. The three registered Student-t seeds remain primary, with seven additional seeds reported separately (Appendix E). Their historical primary criteria were positive energy gain in every seed, an interval excluding zero in at least two seeds, median interval-score gain of at least three percent and at least five of nine domain wins per seed.

For the matched final-calibration comparison we reuse the 36 frozen checkpoints, without refitting. Each held-out series supplies 40 earlier origins for signed joint calibration and 12 later evaluation origins. All calibration targets end no later than the first evaluation cutoff. At each horizon, the correction is the 33rd of 40 signed joint residuals at $\alpha = .2$, with midpoint collapse if needed. All 153 series and all three seeds are retained. Paired inference averages origins within series and then seeds, and resamples series within each of the nine fixed domains. This is an opened-development comparison, not a new-domain confirmation. Endpoint calibration does not define a new joint law, so energy remains a raw-law metric, not a calibrated-law score.

The native-resolution historical comparison instead uses Appliances Energy and Metro Traffic with nominal 90 percent intervals. For each seven-day test episode, the preceding 28–7, 7–5, 5–3 and 3–0 days supply fitting, internal calibration, selection and final calibration. Exact-time features and one common release queue enforce zero or one extra native tick of assumed delay. Eight fixed-calibration and four online methods are reported on common origins. Ten energy episodes receive approximate adjusted block intervals; seven eligible traffic episodes remain descriptive. The release assumptions are not measured publication logs (Appendix F.2).

M COMPLETE RETAINED FORECASTING RESULTS

M.1 A FITTING GAIN PARTLY SURVIVES FINAL CALIBRATION

The local coefficient improves raw forecasts over its parameter-identical amortized control in all three seeds (Table 38).

The matched final-calibration comparison retains a smaller positive gain (Table 38). Averaged over seeds and series, raw interval-score gain is 5.34 percent $[4.32, 6.42]$, and final gain is 1.52 percent $[0.85, 2.18]$. Final mean scores are 6.6822 versus 6.7851; the primary paired difference is -0.1029 $[-0.1522, -0.0557]$. Widths are 4.9311 versus 5.0488, a ratio of .9767 $[.9686, .9850]$. At the nominal 80 percent level the score adds to width ten times the distance by which the outcome falls outside the interval. The paired contrast decomposes into a width part, -0.1178 $[-0.1640, -0.0735]$, and an outside-the-interval part, $+0.0149$ $[-0.0159, +0.0476]$, which sum to the primary -0.1029 to machine precision. All of the surviving gain is narrower intervals. Before calibration the contrast is carried instead by fewer misses at essentially equal width; the sign of the final outside penalty is unresolved. Coverage is .8120 versus .8146, with difference $-.0026$ $[-.0059, .0007]$. The latter interval is not proof of equal coverage. Two of the three per-seed intervals exclude zero favorably and one does not. Passing the seven further replication fits through the same final calibration gives ten seeds in all, all positive, with mean final gain 3.13 percent and a seed-level interval of $[1.80, 4.46]$; the raw gain over the same ten seeds, recomputed from the saved checkpoints, is 7.05 percent $[5.51, 8.59]$.

An outcome-dependent equal-coverage standardization retains advantage 0.0912 against the original 0.1029; it is descriptive, not deployable or causal. Dropping any single domain leaves the pooled gain between 1.05 and 1.91 percent. Calibration attenuates this fitting gain without eliminating it, on this surface and at this unit. A post hoc width/outside split is descriptive and does not establish a universal attenuation rule.

Energy evaluates the original path law: the pooled gain is 3.55 percent $[2.79, 4.34]$. We do not report a fictitious calibrated-law energy score from adjusted interval endpoints. No evaluation label selected a new head or calibration pool, but all nine domains had informed development.

M.2 A PRESPECIFIED EXTERNAL TEST OF ATTENUATION

We fixed the raw-to-final comparison before fitting on two previously unscored groups of seven datasets each. The architecture, three seeds and training procedure are fixed; each dataset is held out from fitting, and its historical calibration precedes evaluation. Table 39 reports equal-dataset contrasts, not the series-weighted development estimand above. The hourly contrast remains favorable and shrinks to 0.6205 of its raw size, with interval $[0.4843, 0.8162]$. The subhourly contrast remains unfavorable; its retained fraction is 0.7414 $[0.2335, 1.1114]$. Thus both point estimates satisfy the specified attenuation criterion, but only the hourly interval lies below one. The width-dominance clause holds on the hourly group and fails on the subhourly group. A corrected, previously examined daily comparison retains 0.9978 of its contrast, almost all of it. Appendix M.6 reports every dataset, both uncertainty calculations, and weighting sensitivities. These historical tests distinguish shrinking an advantage from shrinking a disadvantage; they establish neither universal attenuation nor universal gains.

Table 38: The same frozen fits before and after signed joint calibration. The pooled final-score contrast is primary; per-seed intervals are descriptive series-within-domain bootstrap intervals, and positive gains favor local fitting.

Seed	Raw IS gain (%)	Final IS gain (%)	95% interval	Final width ratio
1	4.53	2.76	[1.81, 3.76]	0.967
2	4.76	0.69	[-0.26, 1.62]	0.988
3	6.72	1.05	[0.22, 1.85]	0.976
Pooled	5.34	1.52	[0.85, 2.18]	0.977

M.3 Fitting and transfer require different evidence

A within-model fitting gain does not establish competitiveness with another model or transfer to another task. Against official-class bounded implementations on identical rows, the history-local fit has median interval-score difference -0.037 $[-0.607, 0.166]$ versus DeepAR (Salinas et al., 2020) and $+1.822$ $[1.431, 2.468]$ versus N-HiTS (Challu et al., 2023), across 153 series; positive differences favor the comparator. Neither result claims hyperparameter parity (Appendix E). Across 538 series held out from this model's development, the separate equal-domain energy difference is $-.0194$ $[-.0301, -.0097]$, but only four of nine declared criteria hold; interval score reverses on Tourism and width increases on FRED-MD (Appendix E.4).

A different 163,594-parameter head on frozen Chronos-2 features fails all eight development criteria, with score and energy worse than the native model. A later native-initialized head passes its development criteria but fails its 293-series transfer comparison. Those historical intervals have documented calibration and target-timing defects; a later emitter-separated, temporally admissible correction retains the negative direction under all six protocols. Both versions, including the favorable development results and all failed transfers, remain in Appendix D.4. The correction does not validate the original interval pipeline.

M.4 Selection versus fixed and online controls

In the full chronological comparison, every tested selection rule has higher aggregate final score than each of the three fixed QR controls in all four source–delay settings (Table 40). In each of the four settings the lowest score belongs to a rule that revises the issued interval online: clipped trailing recalibration on both energy settings, a bootstrap sequential rule on traffic at the shorter delay, and clipped adaptive conformal inference on traffic at the longer delay. In three settings a second online rule is next, while the bootstrap ensemble rule, also online, ranks eleventh or twelfth in all four; the advantage belongs to particular online rules rather than to online revision as such. These rankings concern four aggregates, not 408 independent trials or a claim that each selection rule loses on every episode.

All 40 requested episode–delay cells are accounted for: 34 qualify, yielding 408 method cells, with six exclusions unreplaced. The primary calibrated-loss versus joint-QR energy contrast under delay one is unfavorable across all three specified blocks; its adjusted interval is $[5.39, 94.47]$. Other primary energy intervals include zero, which does not establish equivalence. Traffic has only seven eligible episodes and is descriptive. Under its delayed setting, calibrated selection improves over raw-loss selection by 299.93, in six episodes, but is worse than joint QR by 272.46. The complete coverage, width, raw/final and available-origin results are in Appendix F.2; the target of $.90$ is not the observed coverage.

Native bootstrap SPCI and EnbPI have separate shared-fit, mature-feedback comparisons (Appendix F.1). A native RetroAdj comparison covers 17 eligible windows: rolling joint QR has lower score on energy and RetroAdj on traffic. Different fitting allocations prevent treating this as a calibration-only effect (Appendix F.3). We do not infer native-algorithm superiority from the fixed-center adapters.

M.5 An optimized internal score is not a final-risk result

In the independent synthetic comparison, complete event search and ordinary same-event search emit identical endpoints, as their definitions require. Among 640 scenario–replicate–calibrator cells, no

Table 39: External raw and issued interval-score contrasts, local minus control. Negative values favor local fitting. Retention is $|F|/|R|$; intervals resample series within each fixed dataset. Daily results are corrected secondary evidence.

Frequency	Datasets/series	R	F	Retention [95% interval]
Hourly	7/215	-0.1144	-0.0710	0.6205 [0.4843, 0.8162]
Subhourly	7/394	0.0380	0.0282	0.7414 [0.2335, 1.1114]
Daily (secondary)	4/289	-16.7403	-16.7028	0.9978 [0.9746, 0.9987]

one of the 20 scenario–calibrator comparisons establishes a favorable adjusted final-score difference from both raw selection and no update. In the correctly specified signed-joint case, every one of the 32 internal scores improves, but only 14 final scores improve; the final mean difference is $+.01424$ with adjusted interval $[-.01864, .04711]$. This is not an equivalence conclusion, and the separate positive radius witness does not predict a positive result for this center-action family.

A separate fixed-budget test retains fitted quantile controls and fixes the final pool at nine observations. Two pilot allocation rules show no consistent advantage over ordinary selection on ten synthetic settings and 32 historical episode–delay cells (Appendix H). The strict-ranking example is therefore a boundary result, not evidence for a winning allocation algorithm.

M.6 A SEPARATELY SPECIFIED EXTERNAL COMPARISON

The external comparison asks whether the absolute difference between two models shrinks after final calibration. It does not ask whether the local model always wins. For each dataset we average over its evaluated rows and three seeds, then weight datasets equally. Write the candidate-minus-control contrasts as R before calibration and F afterward. The primary criterion is $|F|/|R| < 1$ on each of the two previously unscored frequency groups. A separate clause asks whether the final width contrast dominates the outside penalty in absolute value, $|W| > |O|$, where $F = W + O$.

Data and roles. The fixed TIME snapshot supplies seven hourly datasets with 215 series and seven subhourly datasets with 394 series. Dataset and frequency eligibility were fixed from metadata before values were used to fit these models. Azure's three datasets share a source; an additional equal-source analysis treats them as one. These are historical held-out-dataset tests, not future observations or verified real-time release protocols. Each leave-one-dataset-out fold fits both 52,986-parameter models on the other datasets of its group, using three fixed seeds. There are 18 folds and 108 fits including the four-dataset, 289-series daily comparison; the latter uses previously examined sources and is secondary. Its 289 series are 36 customer, 51 location, four cryptocurrency and 198 port-activity series under the 80-origin eligibility rule below. This is a different population from the 293-series daily Chronos-2 head-transfer study retained in Appendix D.4; the two denominators are not interchangeable.

Each series contributes 80 origins with context 336, horizon 16 and stride 16. The first 12 origins construct the representation, the next 12 determine its geometry, the next 16 form the remaining pre-evaluation role, and the next 12 are evaluated; the last 28 remain unused. Only targets ending by the first evaluation cutoff enter calibration. Context and targets share the same context-median and robust-scale normalization, with the fixed interpolation and clipping procedure. Signed joint calibration uses 40 admissible origins at $\alpha = .2$. Architecture, optimizer and training schedule are unchanged across the two arms; a held-out dataset contributes calibration but no model fitting. The

Table 40: All twelve methods on common origins, reporting final interval score in Wh for energy and vehicles/hour for traffic. Columns use zero or one assumed extra native tick of delay; online methods receive additional mature feedback. Rows are named by what each rule does; Appendix F maps every row to the published algorithm it implements.

Method	Energy, 0	Energy, 1	Traffic, 0	Traffic, 1
Linear QR	232.65	309.44	2233.45	3345.53
Clipped-feature QR	232.40	309.50	2233.45	3335.31
Joint clipped-feature QR	232.41	309.64	2214.82	3329.81
Joint clipped-feature posterior	340.50	427.16	2852.08	4514.86
Raw-loss selection	235.14	348.13	2441.52	3902.20
Calibrated-loss selection	237.40	348.91	2539.87	3602.27
Complete event search	237.42	349.11	2540.21	3603.66
Ordinary event grid	237.42	349.11	2540.21	3603.66
Adaptive miscoverage update	230.51	309.53	2217.14	3301.91
Trailing residual update	230.35	307.78	2209.83	3303.33
Sequential residual update	285.25	397.21	1943.60	3895.26
Bootstrap residual update	341.65	431.78	2825.18	4262.43

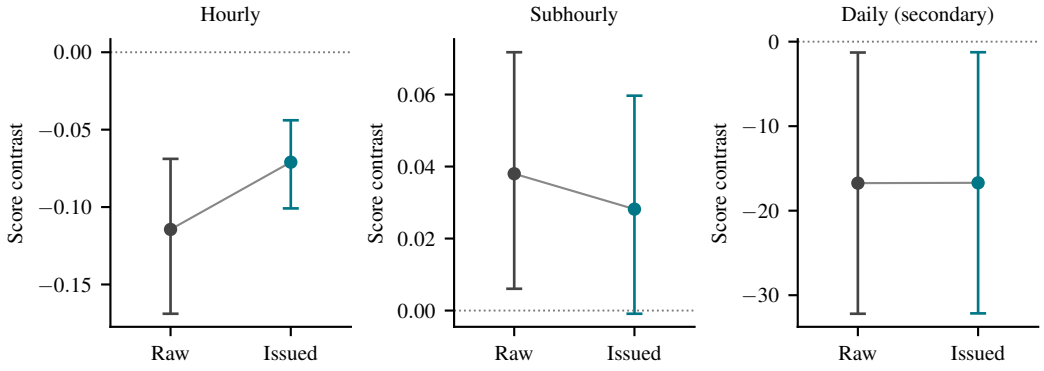

Figure 5: Raw and issued contrasts on all three frequency groups, with the registered series-within-dataset intervals. Negative contrasts favor local fitting. Panels have different vertical scales. The daily comparison is secondary.

Table 41: Complete aggregate uncertainty. Relative gains average dataset-specific percentages; positive values favor local fitting. The seed-mean retention has no corresponding pooled series interval because it is a different estimand.

Frequency	Quantity	Estimate	Series interval	Seed interval
Hourly	Width	-0.0548	[-0.0881, -0.0241]	[-0.2222, 0.1125]
Hourly	Outside	-0.0162	[-0.0375, 0.0054]	[-0.1970, 0.1647]
Hourly	Raw gain (%)	3.6861	[1.9926, 6.0245]	[1.7315, 5.5385]
Hourly	Final gain (%)	2.2854	[1.2771, 3.5387]	[1.0923, 3.3481]
Hourly	Seed-mean retention	0.6235	not the pooled estimand	[0.3292, 0.9177]
Subhourly	Width	-0.0135	[-0.0307, 0.0034]	[-0.0246, -0.0023]
Subhourly	Outside	0.0416	[0.0127, 0.0726]	[-0.0367, 0.1200]
Subhourly	Raw gain (%)	-2.5147	[-3.3505, -1.6752]	[-4.5042, -0.6178]
Subhourly	Final gain (%)	-1.8774	[-2.6850, -1.0918]	[-4.2480, 0.4724]
Subhourly	Seed-mean retention	0.6136	not the pooled estimand	[0.0262, 1.2010]
Daily (secondary)	Width	-16.4610	[-31.8314, -1.0908]	[-67.8657, 34.9437]
Daily (secondary)	Outside	-0.2418	[-0.3791, -0.1045]	[-0.3949, -0.0887]
Daily (secondary)	Raw gain (%)	23.2938	[15.4407, 23.9989]	[5.1797, 34.4418]
Daily (secondary)	Final gain (%)	23.4212	[15.7040, 24.1124]	[5.4164, 34.3920]
Daily (secondary)	Seed-mean retention	0.9888	not the pooled estimand	[0.9480, 1.0297]

hardware amendment preceded the fits. V100 devices with 16 or 32 GB were used under the same scientific configuration.

Uncertainty and decision. The primary 95 percent percentile intervals use 10,000 series resamples within each fixed dataset. Each series retains both models, all origins and all seeds. A one-series dataset contributes no within-dataset resampling variation. These intervals do not sample a population of datasets or refit models. Separate Student intervals over the three seeds have two degrees of freedom. In particular, the seed-mean retention is a mean of three ratios, not the primary ratio of aggregate contrasts. Both primary point estimates meet the prespecified criterion, but only the hourly primary interval is below one. The hourly local model has lower score before and after calibration; the subhourly local model has higher score at both stages. The width clause holds only on the hourly group. Seven hourly datasets and three of seven subhourly datasets have retention below one. These outcomes do not establish a universal attenuation law.

Weighting sensitivities. Merging the three Azure datasets into one source gives subhourly retention 0.7860 $[0.4896, 0.9809]$, but the issued contrast is still unfavorable. This sensitivity does not replace the primary interval containing one. Equal row weighting changes the sign of the subhourly contrast and gives retention 0.9046. On daily data, equal-dataset retention is 0.9978, whereas equal-row

Table 42: Post hoc dataset omissions. Equal weights over the remaining datasets; negative raw (R) and final (F) contrasts favor local fitting.

Group	Omitted dataset	R	F
Hourly	Australia_Solar	-0.08060	-0.03247
Hourly	CPHL	-0.14880	-0.09326
Hourly	Coastal_T_S	-0.10851	-0.06762
Hourly	NE_China_Wind	-0.09692	-0.07347
Hourly	SG_PM25	-0.14960	-0.09276
Hourly	current_velocity	-0.11851	-0.07651
Hourly	epf_electricity_price	-0.09816	-0.06098
Subhourly	EWELD_Load	$+0.01430$	$+0.00230$
Subhourly	Finland_Traffic	-0.01973	-0.01784
Subhourly	OpenElectricity_NEM	$+0.05742$	$+0.04952$
Subhourly	Water_Quality_Darwin	$+0.05910$	$+0.04443$
Subhourly	azure2019_D	$+0.04426$	$+0.03371$
Subhourly	azure2019_I	$+0.04378$	$+0.03174$
Subhourly	azure2019_U	$+0.06688$	$+0.05334$
Daily	Crypto	$+0.01120$	$+0.00463$
Daily	Port_Activity	-22.32694	-22.27082
Daily	Supply_Chain_Customer	-22.32284	-22.27245
Daily	Supply_Chain_Location	-22.32250	-22.27255

Table 43: Every registered weighting. Dataset weighting is primary; row and shared-source weighting are sensitivities.

Frequency	Weighting	R	F	$\lvert F/R\rvert$
Hourly	Dataset	-0.1144	-0.0710	0.6205
Hourly	Row	-0.1104	-0.0564	0.5108
Subhourly	Dataset	0.0380	0.0282	0.7414
Subhourly	Source	0.0707	0.0556	0.7860
Subhourly	Row	-0.0311	-0.0281	0.9046
Daily (secondary)	Dataset	-16.7403	-16.7028	0.9978
Daily (secondary)	Source	-22.3227	-22.2725	0.9978
Daily (secondary)	Row	-0.9117	-0.9221	1.0115

retention is 1.0115. Crypto supplies the daily favorable contrast; each of the other three datasets has a small unfavorable final contrast. These are changes of estimand, not replications of the same effect.

Post hoc dataset omissions. Table 42 recomputes every leave-one-dataset-out point contrast from the saved full-precision aggregates, without refitting. Finland Traffic has one series and therefore no within-dataset bootstrap variation. Omitting it changes both signs in the subhourly comparison; this is sensitivity to the fixed dataset population, not a new confidence interval. No omission replaces a registered primary result.

Corrected and withdrawn daily runs. The earlier daily run used unnormalized, clipped context against normalized targets. Its outputs remain archived but are withdrawn as evidence. The corrected daily run is adjacent secondary evidence, not another previously unscored confirmation, and its nearly unit retained contrast (0.9978) shows almost no attenuation.

The archive retains every reported quantity. Independent reconstruction checks paired score rows and both uncertainty calculations without refitting.

Table 44: All three fitted seeds in each group, with equal dataset weights.

Frequency	Seed	R	F	$\lvert F/R\rvert$	Final gain (%)
Hourly	20260827	-0.1444	-0.0794	0.5497	2.643
Hourly	20260828	-0.1110	-0.0844	0.7601	2.278
Hourly	20260829	-0.0879	-0.0493	0.5606	1.740
Subhourly	20260827	0.0285	0.0140	0.4906	-1.692
Subhourly	20260828	0.0731	0.0648	0.8862	-2.920
Subhourly	20260829	0.0124	0.0057	0.4640	-1.051
Daily (secondary)	20260827	-40.1896	-40.1313	0.9986	24.547
Daily (secondary)	20260828	-1.2492	-1.2115	0.9699	13.358
Daily (secondary)	20260829	-8.7820	-8.7655	0.9981	21.808

Table 45: Every hourly dataset. The final difference satisfies $F = W + O$; tiny nonzero quantities are printed in scientific notation.

Dataset	Series	R	F	$\lvert F/R\rvert$	W	O
Australia Solar	1	-0.3175	-0.3023	0.9521	-0.5344	0.2321
CPHL	2	0.0917	0.0625	0.6816	0.1757	-0.1131
Coastal T S	72	-0.1501	-0.0913	0.6087	-0.0740	-0.0173
NE China Wind	4	-0.2196	-0.0562	0.2562	0.0826	-0.1389
SG PM25	5	0.0965	0.0595	0.6164	0.0479	0.0116
current velocity	126	-0.0900	-0.0380	0.4222	-0.0446	0.0066
epf electricity price	5	-0.2121	-0.1312	0.6184	-0.0371	-0.0941

Dataset	Raw gain (%)	Final gain (%)	Raw cov.	Final cov.
Australia Solar	8.695	9.321	0.783/0.715	0.870/0.905
CPHL	-2.377	-1.635	0.715/0.701	0.752/0.721
Coastal T S	3.091	1.915	0.840/0.844	0.822/0.823
NE China Wind	10.427	2.809	0.825/0.724	0.882/0.842
SG PM25	-2.274	-1.423	0.882/0.861	0.824/0.816
current velocity	2.478	1.084	0.819/0.833	0.797/0.800
epf electricity price	5.763	3.926	0.917/0.927	0.878/0.858

Table 46: Every subhourly dataset. Coverage entries give local/control; nominal coverage is .8. Ratio instability near zero raw differences remains visible.

Dataset	Series	R	F	$\lvert F/R\rvert$	W	O
EWELD Load	10	0.1802	0.1834	1.0176	-0.0208	0.2042
Finland Traffic	1	0.3844	0.3043	0.7915	0.0715	0.2327
OpenElectricity NEM	10	-0.0786	-0.0999	1.2723	-0.0330	-0.0670
Water Quality Darwin	36	-0.0886	-0.0694	0.7833	-0.0441	-0.0253
azure2019 D	128	0.0005	-0.0051	11.2171	0.0209	-0.0260
azure2019 I	128	0.0033	0.0068	2.0279	-0.0190	0.0258
azure2019 U	81	-0.1353	-0.1228	0.9079	-0.0699	-0.0529

Dataset	Raw gain (%)	Final gain (%)	Raw cov.	Final cov.
EWELD Load	-4.952	-5.113	0.794/0.820	0.790/0.798
Finland Traffic	-19.645	-15.439	0.733/0.814	0.819/0.856
OpenElectricity NEM	3.406	4.352	0.756/0.768	0.803/0.795
Water Quality Darwin	1.969	1.560	0.672/0.670	0.796/0.792
azure2019 D	-0.008	0.086	0.805/0.799	0.798/0.796
azure2019 I	-0.058	-0.118	0.822/0.822	0.801/0.801
azure2019 U	1.685	1.530	0.832/0.832	0.812/0.809

Table 47: Every corrected daily dataset, with the same columns and secondary status.

Dataset	Series	R	F	$\lvert F/R\rvert$	W	O
Crypto	4	-66.9947	-66.8251	0.9975	-65.8368	-0.9883
Port Activity	198	0.0197	0.0013	0.0637	-0.0074	0.0086
Supply Chain Customer	36	0.0074	0.0062	0.8297	0.0118	-0.0057
Supply Chain Location	51	0.0064	0.0065	1.0065	-0.0116	0.0180

Dataset	Raw gain (%)	Final gain (%)	Raw cov.	Final cov.
Crypto	93.751	93.971	0.681/0.622	0.756/0.737
Port Activity	-0.281	-0.019	0.654/0.655	0.798/0.797
Supply Chain Customer	-0.163	-0.135	0.810/0.809	0.768/0.768
Supply Chain Location	-0.131	-0.132	0.767/0.771	0.770/0.771

N Fixed-Level Calibration and the Cost of Assessment

Fixing the requested coverage does not by itself make large calibration pools easy to compare. We first distinguish final-pool variation from assessment error, then give a finite-order variance calculation and an explicit fixed-level information obstruction. All predictors below are fixed independently of the observations used for selection.

N.1 The finite-pool projection

For signed calibration of $[z(x) - \nu c(x), z(x) + \nu c(x)]$, with constant $\nu \geq 0$, put $E = |Y - z(X)|/C$ and $C = c(X) > 0$. The issued halfwidth is CQ_m, where $Q_m = E_{(k)}$, $k = \lceil (m+1)(1-\alpha) \rceil \leq m$. Define

$$\ell(t; e, c) = 2ct + \frac{2c}{\alpha}(e - t)_+, \qquad G(t) = \mathbb{E}\ell(t; E, C), \qquad R_m = \mathbb{E}G(Q_m).$$

At a non-atomic q with $\Pr(E > q) = \alpha$, differentiation gives

$$G'(q) = 2\mathbb{E}C - \frac{2}{\alpha}\mathbb{E}[C\mathbf{1}\{E > q\}] = -\frac{2}{\alpha}\operatorname{Cov}(C, \mathbf{1}\{E > q\}). \tag{20}$$

Thus a regular residual density alone does not remove first-order sensitivity to the final calibration quantile. Width/coverage independence already motivates orthogonal quantile regression (Feldman et al., 2021); quantile-nuisance orthogonality also appears in expected-shortfall inference (He et al., 2022). The identity above is not a new training or inference principle.

For finite support, let F be the residual CDF and $H(u) = \Pr\{\mathrm{Bin}(m, u) \geq k\}$, $b(u) = H'(u)$. Tail integration and differentiation of the contaminated law give

$$R_m = G(0) + \int_0^\infty G'(t)[1 - H(F(t))]dt, \tag{21}$$

$$\psi(e, c) = \mathbb{E}_{Q_m}\ell(Q_m; e, c) - R_m - \int_0^\infty G'(t)b(F(t))[\mathbf{1}\{e \leq t\} - F(t)]dt. \tag{22}$$

Here $b(u) = m\binom{m-1}{k-1}u^{k-1}(1-u)^{m-k}$. Atoms do not invalidate the integral: its integrand is evaluated between the finitely many breakpoints. A contamination point outside the original support enlarges the integration range. The first term changes the query law; the second changes the calibration law. Differentiating the symmetric $d = m + 1$ row representation independently yields

$$\psi(z) = d\{\mathbb{E}[K(z, Z_2, \ldots, Z_d)] - R_m\}.$$

For a pair, the relevant projection is $\psi_a(Z) - \psi_b(Z)$ on the same observation, not the difference of two independent copies. These population identities contain unknown distributions and are not confidence procedures.

N.2 An explicit finite-order remainder

Let $h = K_a - K_b$ be square integrable and symmetric, and let v_j denote the variance of its canonical order-j Hoeffding projection. The classical orthogonal decomposition (Maurer, 2022) gives

$$\operatorname{Var}(h) = \sum_{j=1}^{d}\binom{d}{j}v_j, \qquad \operatorname{Var}(U_N) = \sum_{j=1}^{d}\frac{\binom{d}{j}^2}{\binom{N}{j}}v_j.$$

For every $N \geq d$, therefore,

$$0 \leq \operatorname{Var}(U_N) - \frac{d^2 v_1}{N} \leq \frac{d(d-1)}{N(N-1)}\{\operatorname{Var}(h) - dv_1\}. \tag{23}$$

Indeed, $\binom{d}{j}/\binom{N}{j}$ is nonincreasing in j, so every term with $j \geq 2$ has at most the order-two factor. This is a classical consequence, not a new general U bound; growing-order processes have substantially richer theory (Minsker & Yao, 2026). For $|h| \leq B$, replacing the braces by B^2 is valid but coarse. When $d^2 v_1$ is of order m and B is bounded, $m/N \to 0$ suffices for this remainder bound to be

negligible relative to the leading term. If $d^2 v_1$ is of order one, the same generic argument needs $m^2/N \to 0$; this sufficient condition is not asserted necessary.

The known global-quantile bound below also applies at $N = d$. Writing its numerator as K_m^2 gives $\mathrm{Var}(h) \le K_m^2/(2d)$, hence the remainder in equation 23 is at most $(d-1)K_m^2/[2N(N-1)]$. This $O(m/N^2)$ consequence for fixed class constants combines two existing bounds; it is not an improvement over the same generic expression using the actual kernel variance and projection. Zero calibration sensitivity also need not remove the query term: for $E \sim \mathrm{Uniform}[0,1]$, $C = 1$ and $\alpha = .1$, $G'(.9) = 0$, whereas the limiting influence $20(E - .9)_+ - .1$ has variance $37/300$. This limiting calculation is not the finite-m variance of the complete estimate.

N.3 A FIXED-LEVEL INFORMATION OBSTRUCTION

The next construction shows that the leading coefficient really can grow with m, even though α is fixed and the paired kernel is bounded.

Proposition 9 (Fixed-level calibration-order cost)**.** *Fix $\alpha = 1/2$, odd $m \ge 1$, $d = m + 1$, and $0 < \epsilon \le 1/4$. Let $X \sim \mathrm{Bernoulli}(p)$, $C(0) = 1$, $C(1) = 1/3$ and $Y = C(X)X$. The reference center is zero, and the candidate center is $\epsilon C(X)u(X)$, where $u(0) = 1, u(1) = 1/3$. Both have raw halfwidth $\nu C(X)$ for any fixed $\nu \ge 0$, and use signed calibration. Set $T_m(p) = \Pr\{\mathrm{Bin}(m,p) \ge d/2\}$. Their exact fresh risk difference is*

$$D_m(p) = 2\epsilon \left[1 - \frac{14p}{9} + \left(-\frac{4}{3} + \frac{16p}{9}\right) T_m(p)\right]. \tag{24}$$

At $p = 1/2$, $D_m = 0$, and with $t_m = m\binom{m-1}{(m-1)/2}/2^{m-1}$,

$$d^2 v_1 = \frac{\epsilon^2}{4}\left(\frac{4}{3} + \frac{8}{9} t_m\right)^2 \ge \frac{8m\epsilon^2}{81}. \tag{25}$$

Consequently $\mathrm{Var}(U_N) \ge 8m\epsilon^2/(81N)$ for $N \ge d$. Moreover, for $0 < \gamma \le \epsilon/100$ and $0 < \delta < 1/2$ there are two such laws, with the same fixed candidates, whose contrasts are respectively at most $-\gamma$ and at least γ. Any selector choosing the better action with probability at least $1 - \delta$ under both laws requires

$$N \ge \frac{m\epsilon^2}{729\gamma^2}\,\mathrm{kl}(1-\delta, \delta). \tag{26}$$

Here N counts every accessible independent observation informative about p, including covariates and any fitting observations. The final pool is unopened during selection. This is not a label-only lower bound if unlimited informative covariates are free.

Proof. The baseline residuals are $0, 1$ and the candidate residuals are $\epsilon, 1 - \epsilon/3$, in the same order. A size-m calibration sample uses the high residual precisely when it contains at least $d/2$ ones. On the low event the paired query mean is $2\epsilon(1 - 14p/9)$; on the high event it is $-(2\epsilon/3)(1 - 2p/3)$. Averaging proves equation 24. Symmetry gives $T_m(1/2) = 1/2$, and

$$D_m'(1/2) = -4\epsilon/3 - (8\epsilon/9)t_m.$$

For a Bernoulli law the influence function is $D_m'(p)(X - p)$, so $d^2 v_1 = p(1-p)D_m'(p)^2$. The recurrence $t_1 = 1$, $t_{m+2}/t_m = (m+2)/(m+1)$ proves by induction $(m+1)/2 \le t_m^2 \le m$. This gives equation 25; equation 23 gives the variance lower bound.

For the testing claim set $h_0 = 9\gamma/(2\epsilon t_m)$ and $p_\pm = 1/2 \pm h_0$. The gap restriction ensures $h_0 \le 1/(8\sqrt{m})$. On this neighborhood,

$$T_m'(p) = t_m\{1 - 4(p - 1/2)^2\}^{(m-1)/2} \in [t_m/2, t_m], \quad T_m(p) \le 5/8, \quad p \le 5/8.$$

For $m = 1$ the derivative is constant; for odd $m \ge 3$ the lower bound follows from Bernoulli's inequality. Differentiating the bracket in equation 24 bounds it by $-t_m/9$, whence $D_m'(p) \le -2\epsilon t_m/9$. Integration from $1/2$ gives opposite gaps of at least γ at $p_\pm$. The one-row divergence is bounded by its chi-square divergence,

$$\mathrm{KL}(P_+ \| P_-) \le \frac{4h_0^2}{p_-(1-p_-)} \le 18h_0^2.$$

Because X determines (C, Y), this is the divergence of the full observation. Data processing for the output action yields $18Nh_0^2 \geq \mathrm{kl}(1-\delta, \delta)$. Substitute h_0 and $t_m^2 \geq m/2$ to obtain equation 26. Randomization independent of the world cannot increase the information available to the rule. □

The construction is atomic and uses scale–tail dependence. It establishes neither the same rate at every fixed α nor a smooth-density minimax result. The elementary two-point testing tool is classical; the specific increment is its finite-calibration realization at a fixed level. For the same centers with $C \equiv 1$, the coefficient at $p = 1/2$ instead equals $16\epsilon^2/9$ for every odd m. This comparison concerns projection sensitivity, not matched risk gaps or an algorithmic performance advantage.

N.4 Exact checks and comparisons

A separate result-before-execution protocol specifies $\alpha = 1/2$ with $m \in \{1, 3, 9, 19, 39\}$ and $\alpha = .1$ with $m \in \{9, 19, 39\}$; $C(1) \in \{1/3, 1\}$; $p \in \{.01, .35, .5, .65, .99\}$; and $N \in \{d, 2d, 8d\}$. These are 240 exact finite-law settings, not Monte Carlo repetitions or additional real datasets. Every risk and variance is a rational probability sum. If a d-block contains r ones, write its paired kernel as h_r. Given T ones among the N observations,

$$U_N(T) = \sum_{r=0}^{d} \frac{\binom{T}{r}\binom{N-T}{d-r}}{\binom{N}{d}} h_r.$$

A binomial average over T gives the exact variance. Independent literal interval construction checks the risk and derivative; a separate overlap- covariance calculation agrees with all 240 variances within 2.78×10^{-17}. Forty small settings are additionally enumerated over every ordered sample and subset. Eighteen rational contamination identities include atoms, finite rank boundaries and points outside support. These checks corroborate the implementation, not the proof's quantifiers.

m	$d^2 v_1/\epsilon^2$	$\mathrm{Var}(U_N)/\epsilon^2$	Remainder fraction	Block/full variance
1	1.2346	0.0837	0.0786	2.1007
3	1.7778	0.0592	0.0609	2.0087
9	3.0991	0.0403	0.0384	1.9010
19	4.9857	0.0320	0.0253	1.8303
39	8.3838	0.0266	0.0161	1.7697

Table 48: Exact selection-sample variance at $\alpha = p = 1/2$, $N = 8(m+1)$, $C(1) = 1/3$ and $\epsilon = 1/4$. The remainder fraction is $(\mathrm{Var}(U_N) - d^2 v_1/N)/\mathrm{Var}(U_N)$. These are analytic expectations, not estimated error bars.

For independent blocks the variance is $\mathrm{Var}(h)/\lfloor N/d \rfloor$. For M uniform subsets sampled independently with replacement, conditional on the full sample, the incomplete average has unconditional variance

$$\mathrm{Var}(U_N) + \{\mathrm{Var}(h) - \mathrm{Var}(U_N)\}/M.$$

This includes subset overlap and conditional Monte Carlo variation. We use $M = N$; none of these variance comparisons is an equal-walltime superiority claim. Complete tables and code accompany the paper.

For a distinct smooth check, $E \sim \mathrm{Uniform}[0, 1]$ and $C = 1 + \kappa E$, $\kappa \in \{0, 1\}$ give $G'(1-\alpha) = -\kappa(1-\alpha)$. Exact beta moments reproduce the conditional-risk standard deviations $(.013666, .042934)$, $(.001283, .009139)$ and $(.000127, .002719)$ at $m = 99, 999, 9999$ and $\alpha = .1$. These describe $G(Q_m)$, not U_N. Smooth quantile expansions and allocation already have direct prior theory (Das et al., 2026). The analytic laws here do not supply an estimable variance certificate for unknown populations or establish useful fitted updates; the five fitted rules in the main comparison still make zero updates.

N.5 A SHARED-SCALE LABEL BOUND AT NINETY PERCENT COVERAGE

The preceding example charges informative covariates. The next one holds their distribution fixed, so even unlimited unlabeled inputs cannot remove the need for labels. Both predictors also stay fixed as the calibration sample grows.

Proposition 10 (Fixed-level label complexity)**.** *Fix $\alpha = 1/10$, $m = 10s - 1$ for integer $s \geq 1$, and $h > 0$. Let $X \sim \mathrm{Bernoulli}(1/2)$ in both worlds. Conditional on $X = 0$, let $Y = 0$; conditional on $X = 1$, let $Y = 2$ with probability $2p$ and otherwise zero. Both predictors use $C(X) = 1 + X$ and zero raw halfwidth. Their centers are $z_A = 0$ and $z_B = cC(X)$, with the fixed $c = 7/6$. Use signed or nonnegative joint calibration at rank $k = 9s$. For $p_\pm = .1 \pm h/\sqrt{m}$ and all sufficiently large m, B has smaller fresh risk under p_- and A under p_+. Both gaps tend to $\Delta_h = \Phi(h/.3) - 1/2 > 0$.*

Any procedure observing at most N informative labels, possibly choosing which independent rows to label using arbitrary unlabeled covariates and earlier labels, and selecting the better predictor with probability at least $1 - \delta$ in both worlds, satisfies

$$N \geq \frac{m}{100h^2}\mathrm{kl}(1-\delta, \delta), \qquad 0 < \delta < 1/2, \tag{27}$$

provided $h/\sqrt{m} \leq 1/20$ and the stated risk ordering holds. Every accessible fitting, pilot or preliminary calibration label counts toward N; the final calibration pool and query are independent and hidden until after selection. For fixed h, ordinary independent-block comparison uses $O(m \log(1/\delta))$ passive labels. Thus the order is $\Theta(m \log(1/\delta))$ on this specified family for sufficiently large m and $0 < \delta \leq 1/4$.

Proof. At zero raw halfwidth, standardized residuals are nonnegative, so truncating the calibration adjustment at zero changes nothing. Both operators issue $[z - CE_{(k)}, z + CE_{(k)}]$, where $E = |Y - z|/C$. Write $B_m(p) = \Pr\{\mathrm{Bin}(m, p) \geq s\}$ and $T_m(p) = \Pr\{\mathrm{Bin}(m, p) \geq 9s\}$. Predictor A has standardized absolute residuals zero and one. Its calibration quantile is one exactly on the event defining B_m. The conditional query scores at quantiles zero and one have means $40p$ and 3. Predictor B has residuals c when $Y = 0$ and $c - 1$ when $Y = 2$. Its quantile is $c - 1$ exactly on the event defining T_m. At quantile c, all outcomes are covered and mean width is $3c$. At quantile $c - 1$, the issued interval is $[C, (2c-1)C]$: $Y = 2$ is on its boundary and $Y = 0$ incurs mean distance $3/2 - 2p$. Hence

$$R_A(p) = 40p + (3 - 40p)B_m(p), \tag{28}$$

$$R_B(p) = 3c + (27 - 40p)T_m(p). \tag{29}$$

The triangular-array binomial central limit theorem gives $B_m(p_\pm) \to \Phi(\pm h/.3)$; its bounded summands satisfy Lindeberg's condition since $p_\pm \to .1$. A binomial tail bound gives $T_m(p_\pm) \to 0$ because $k/m \to .9$. With $3c = 3.5$, $R_A - R_B$ therefore tends to $\mp\Delta_h$. In particular, the absolute gaps eventually exceed $\gamma_h = \Delta_h/2$.

Unlabeled X has exactly the same law in both worlds. A queried label at $X = 0$ is uninformative. At $X = 1$ its conditional success probabilities are $q_\pm = .2 \pm 2h/\sqrt{m}$. Under $h/\sqrt{m} \leq 1/20$, $q_+ \in [.2, .3]$, so

$$\mathrm{kl}(q_-, q_+) \leq \frac{(q_- - q_+)^2}{q_+(1 - q_+)} \leq \frac{100h^2}{m}.$$

The adaptive transcript chain rule bounds total divergence by $100Nh^2/m$. Randomized query choices based only on the existing transcript add no further information. Data processing for the final binary action gives equation 27.

For the upper bound, both predictors share the bounded scale $C \leq 2$, and their endpoints differ by cC. Their symmetric paired block kernel has absolute value at most $B = 16c = 56/3$ by Proposition 2; no response truncation is used. The mean of n independent blocks of size $m + 1$ has the wrong sign with probability at most $\exp\{-n\gamma_h^2/(2B^2)\}$. Thus $n \geq 2B^2 \log(1/\delta)/\gamma_h^2$ suffices. This is ordinary Hoeffding comparison, not a new algorithm. The constants are not asserted optimal. □

The specified fixed candidates do not require centering at finite m. For a diagnostic centered variant, set

$$c_m = \{4 - B_m(.1) - 23T_m(.1)\}/3.$$

Then $R_A(.1) = R_B(.1)$ exactly. Reflecting a binomial term $j \geq 9s$ to $m - j \leq s - 1$ gives $T_m(.1) \leq 9^{-(8s+1)}[1 - B_m(.1)]$, so $1 < c_m < 4/3$ for every s; also $c_m \to 7/6$. This choice depends only on the declared construction, not observed labels. It is not needed for the asymptotic theorem.

Finite checks retain the small-sample boundary. A separate frozen calculation uses $m \in \{9, 49, 169, 529, 1849\}$, $h \in \{.1, .2\}$, both worlds and both center choices: forty exact rational risk points. Independent literal interval construction and an 80-digit recurrence agree within 2.57×10^{-77}; all eight $m = 9$ points additionally use exact three-state multinomial enumeration. The calculation used signed unit halfwidth, whose intervals equal those of either operator at zero halfwidth: $(E - 1)_{(k)} + 1 = E_{(k)} \geq 0$. No risk values are refitted or recomputed for this identity. Nonnegative calibration at unit halfwidth is a different procedure.

		$c = 7/6$		$c = c_m$	
m	h	p_-	p_+	p_-	p_+
9	0.1	-0.67915	+0.14364	-0.56657	+0.25622
9	0.2	-1.72840	+0.21062	-1.61582	+0.32320
49	0.1	-0.24793	+0.01444	-0.19762	+0.06475
49	0.2	-0.60416	-0.01156	-0.55385	+0.03875
169	0.1	-0.07934	-0.04674	-0.05211	-0.01951
169	0.2	-0.21678	-0.11284	-0.18955	-0.08561
529	0.1	+0.01018	-0.08149	+0.02559	-0.06609
529	0.2	-0.01678	-0.16925	-0.00137	-0.15384
1849	0.1	+0.06561	-0.10376	+0.07385	-0.09552
1849	0.2	+0.10552	-0.20497	+0.11376	-0.19672

Table 49: Exact $R_A - R_B$ in both worlds, at fixed $\alpha = .1$. Negative values favor A. Small calibration pools can reverse the asymptotic ordering or favor the same candidate in both worlds.

The theorem extends the label-cost obstruction to a fixed coverage level within the same shared-scale geometry, and ordinary comparison attains its order. It does not establish an optimal envelope constant, a smooth-density result, a fitted-model improvement or a useful variance-adaptive certificate. The central limit theorem, two-point information argument and block bound are classical (Tsybakov, 2009). Growing-order U-process deviation theory is also prior work (Minsker & Yao, 2026); this construction identifies the label cost of the specified calibrated-score comparison. This decision lower bound is not the conditional variation of one final calibration pool or a variance bound for U_N.

N.6 A SUFFICIENT REGULARITY CONDITION FOR ASSESSMENT

The atomic obstruction is not universal. The following calculation explains a condition under which neighboring calibration ranks are stable enough to avoid an extra order factor in assessment variance. It applies classical replace-one variance control (Efron & Stein, 1981); stability-based risk concentration is also established prior work (Abou-Moustafa & Szepesvári, 2019). This is a sufficient-condition calculation, not a new general concentration inequality.

Proposition 11 (Regular residual quantiles)**.** *Condition on two independently fitted, fixed candidates. Let h be their symmetric paired query/calibration kernel of order $d = m + 1$ and U_N its complete average on $N \geq d$ IID observations from the final-risk population. Assume both individual fresh risks are finite. For each candidate $t \in \{a, 0\}$, suppose the marginal calibration-score quantile function on $[0, 1]$, including finite endpoint limits, is globally L_t-Lipschitz. Suppose each ordered interval score is L-Lipschitz in its calibration adjustment, uniformly over queries. Suppose further that the paired* query *loss is bounded in absolute value by B_q, for every admissible common pool and query. With deterministic class constants, put*

$$K_m = 2B_q + L(L_a + L_0)\sqrt{\frac{2m}{m+1}}.$$

Then, for every integer $1 \le k \le m$,

$$\mathrm{Var}(U_N) \le \frac{K_m^2}{2N} \le \frac{\{2B_q + \sqrt{2}L(L_a + L_0)\}^2}{2N}. \tag{30}$$

This bounds the full statistic, including its higher-order remainder.

Every variance here is conditional on the fitted candidates. Random independent fitting adds $\mathrm{Var}(\mathbb{E}[U_N \mid \text{fit}])$ to the unconditional variance; it is not removed by independence from selection.

Proof. Fix a query and remove the calibration observation to be replaced. Among the remaining $m-1$ scores, insertion of either its old or new value places the kth order statistic between the $(k-1)$th and kth values. At the extremes use the support endpoints. This includes $m = 1$. Couple the scores through their quantile functions and uniform variables. The corresponding uniform spacing has second moment $2/[m(m+1)]$, so the candidate-specific bracket width W_t satisfies

$$\|W_t\|_2 \le L_t\sqrt{2/[m(m+1)]}.$$

Replacing one row of h changes one query term by at most $2B_q$. In each of the other m query terms, only the two calibration adjustments change. The L_2 triangle inequality, with no independence assumption between those two scores or the query terms, gives

$$\|h - h^{(i)}\|_2 \le \frac{2B_q}{d} + \frac{m}{d}L(L_a + L_0)\sqrt{\frac{2}{m(m+1)}} = \frac{K_m}{d}.$$

Exactly the fraction d/N of all subsets contains the replaced observation. Applying the same triangle inequality to their average gives $\|U_N - U_N^{(i)}\|_2 \le K_m/N$. Efron–Stein now yields $\mathrm{Var}(U_N) \le \frac{1}{2}\sum_i \mathbb{E}(U_N - U_N^{(i)})^2 \le K_m^2/(2N)$. □

For common scales bounded by $C_{\max}$, the safe constant $L = 2C_{\max}/\alpha$ applies to the symmetric signed, collapsed and nonnegative operators. Global endpoint proximity supplies the coarse query bound $B_q = 8\epsilon C_{\max}/\alpha$: both final endpoints differ by at most $2\epsilon C$, and interval score is $2/\alpha$-Lipschitz in endpoint ℓ_1 distance. The smaller rank-improved bound on the *averaged* kernel is not a valid substitute for this query bound without a separate proof.

A density bounded below throughout a bounded support interval suffices for the quantile assumption. A positive density only near one quantile does not. Atoms themselves are not the dividing line: the function $q(u) = \max\{0, 2u-1\}$ is 2-Lipschitz although its law has mass $1/2$ at zero. The two-point laws above instead have quantile jumps. The class is nonempty for genuine paired predictors: for $Y \sim \mathrm{Uniform}[0,1]$, $C = 1$, zero raw halfwidth and centers $1/2, 3/5$, the residual quantile functions have global constants $1/2, 1$.

If these constants are genuinely known, Cantelli's inequality gives a one-sided radius $r = K_m\sqrt{(1-\delta)/(2N\delta)}$. Choosing the alternative only when $U_N + r < 0$ then bounds the probability of choosing a nonbeneficial alternative by δ; a fixed allocation handles a finite family. This ordinary implication has not demonstrated nonzero useful updates or equal-compute superiority. It retains $N \ge m+1$, has polynomial confidence dependence, and does not justify plugging in sample-estimated density bounds. The local extension below allows unbounded residuals, but sharper paired constants and an unknown-population practical certificate remain open.

Separate exact implementation checks cover insertion brackets, score Lipschitzness and classical spacing moments. Independent polynomial integration checks small moments, and literal query/calibration enumeration checks the subset replacement identity. These finite checks are not the proof above; no old fitted family or risk experiment was repeated.

N.7 Local quantiles and an explicit boundary cost

The same argument can use a probability window instead of bounded residual support. This applies classical order-statistic spacing and stability tools (Boucheron & Thomas, 2012; Abou-Moustafa & Szepesvári, 2019), not a new concentration principle. Retain the independent-fit, IID, finite-risk,

L, B_q and $N \geq m+1$ assumptions above. In particular, B_q bounds every paired query loss over every legal common pool, not merely the averaged kernel. Suppose both residual quantile functions are locally Lipschitz on the same $[a,b] \subset (0,1)$, with known constants $L_a^{\mathrm{loc}}, L_0^{\mathrm{loc}}$. For $m \geq 3$ and $2 \leq k \leq m-1$, define

$$\pi_{m,k} = \min\big\{1, 2[\Pr\{\mathrm{Bin}(m-1,a) \geq k-1\} + \Pr\{\mathrm{Bin}(m-1,b) \leq k-1\}]\big\}, \tag{31}$$

$$J_m = \min\left\{2B_q, \sqrt{\frac{2L^2(L_a^{\mathrm{loc}} + L_0^{\mathrm{loc}})^2}{m(m+1)} + 4B_q^2\pi_{m,k}}\right\}. \tag{32}$$

Then the complete statistic satisfies

$$\mathrm{Var}(U_N) \leq \frac{(2B_q + mJ_m)^2}{2N}. \tag{33}$$

To verify the bound, represent each marginal residual by a generalized inverse CDF applied to a uniform variable. Randomized probability integral transforms allow atoms; the two candidates' uniforms on one row may remain arbitrarily dependent. Removing the replaced calibration row leaves $m-1$ uniform ranks. The bracket $[U_{(k-1)}, U_{(k)}]$ lies in $[a,b]$ unless one of the binomial events in equation 31 occurs. The factor two is a union bound across candidates, not an independence assumption. On the common good event the paired change is bounded by $L(L_a^{\mathrm{loc}} D_a + L_0^{\mathrm{loc}} D_0)$, where each spacing has second moment $2/[m(m+1)]$. On its complement the paired change is at most $2B_q$. The events are disjoint, so their second moments add; Minkowski bounds the good-event norm. This gives the square root in equation 32, while $2B_q$ is a separately valid global cap. Replacing a block row changes one query term and m calibration terms. The replacement norm is therefore at most $(2B_q + mJ_m)/d$. The fraction d/N of affected subsets and Efron–Stein yield equation 33, without treating overlapping subsets as independent.

For fixed $a < 1-\alpha < b$ and fixed class constants, the bad-event probability decays exponentially as m grows. This recovers an asymptotic $O(1/N)$ bound, but the finite-m term cannot be discarded. At $\alpha = .1, m = 9$ the rank is $k = m$, outside this statement. At $m = 19$, the rank is interior but the bad-event upper bound can equal one. The bound gives zero for identical candidates when their exact paired bound is $B_q = 0$. For nearby candidates the cap ensures vanishing radius with their distance at fixed m, but it may retain an mB_q cost. It does not prove a useful joint-cancellation improvement.

Finite-scale comparison. Fix zero raw halfwidths, unit scales, centers 0 and .1, and standard normal, Student-t_3 or Laplace noise. All have finite risks and unbounded residuals. We use the known window $[.8, .95]$, $\alpha = .1$, $\delta = .05$ and safe constants $L = 20, B_q = 8$. For a symmetric unimodal density f, put $q_0 = F^{-1}(.975)$. The folded residual quantile at center .1 is at most $q_0 + .1$ throughout this window, giving valid local constants $1/[2f(q_0)]$ and $1/[2f(q_0 + .2)]$. For Laplace the exact quantile derivative instead gives both constants 20. These are specified population classes, not estimated density certificates.

The fixed calculation crosses $m \in \{9, 19, 49, 99, 199, 499, 999, 4999\}$, $N/(m+1) \in \{32, 128, 512\}$ and the three laws. Nine of the 72 cases have excluded extreme ranks. No remaining case gives a smaller one-sided Cantelli radius than the ordinary rank-range Hoeffding bound at the same N, δ. For normal noise and $N/(m+1) = 512$, the former is 9.7468 at $m = 19$ and 1.2006 at $m = 4999$; the latter is .08654 in both cases. The symmetric-kernel range is only .8. Thus this sufficient local bound supplies no nonzero controlled update in these cases. This is a finite-bound comparison, not another fitted experiment or a claim that every possible local bound fails.

Robust estimation does not supply missing constants. Higher-order robust U estimation with finite variance is established (Joly & Lugosi, 2016); its explicit order-dependent constants are not an automatic improvement for this kernel. Recent incomplete-U median results (Hieu & Ledent, 2026) concern pairwise kernels, whereas ours has order $m+1$. The sharp mean-estimation moment condition in Minsker (2023) is not a theorem for arbitrary growing-order calibration kernels. Finite but unknown variance alone supplies no everywhere-finite uniform upper confidence bound over all finite-variance laws: a rare mass at an arbitrarily large finite value leaves an all-zero sample likely while its mean is arbitrarily large. Known scale bounds or additional structure change that argument. Our declared geometry already supplies a bounded paired kernel; the unresolved issue is a tighter available radius, not replacing one robust estimator name with another.

N.8 A SUFFICIENT CONDITION FOR PAIRED CANCELLATION

Marginal regularity controls each candidate separately. A more demanding *joint* condition can control their difference directly. Condition on the fits as above. Suppose the two residuals on every row admit the same uniform rank V, so $r_a = Q_a(V)$ and $r_0 = Q_0(V)$. This common-order assumption is stronger than having smooth marginal CDFs. On $[a, b]$, assume absolutely continuous quantiles, positive issued halfwidths, a common scale $c(X) \le C_{\max}$, and known bounds

$$|Q'_a - Q'_0| \le \eta, \qquad |Q'_0| \le L_Q, \qquad \Pr_Z\{M_a(t, Z) \ne M_0(t, Z)\} \le \kappa,$$

where $M_j(t, Z)$ indicates that the query is missed after correction $Q_j(t)$. The probability bound holds throughout the window and derivative bounds almost everywhere. Retain the global paired-query bound B_q and finite risks.

For the paired query loss $H_Z(t)$, differentiation away from hinge crossings gives

$$H'_Z(t) = 2c(X)\left[(Q'_a - Q'_0)(1 - M_a/\alpha) + (Q'_0/\alpha)(M_0 - M_a)\right].$$

Thus $\|H'_Z(t)\|_2 \le D$, where

$$D = 2C_{\max}\left[\eta \max\{1, \alpha^{-1} - 1\} + L_Q\sqrt{\kappa}/\alpha\right].$$

Absolute continuity and Minkowski give $\|H_Z(t) - H_Z(s)\|_2 \le D|t - s|$. Because the residual ranks coincide, one bracket suffices. Define π_0 by equation 31 without its factor two and put

$$J_m^{\text{pair}} = \min\left\{2B_q, \sqrt{2D^2/[m(m+1)] + 4B_q^2\pi_0}\right\}.$$

The independent query permits the derivative bound to be integrated over the calibration bracket. The previous role-count proof then yields $\mathrm{Var}(U_N \mid \text{fit}) \le (2B_q + mJ_m^{\text{pair}})^2/(2N)$. Global conditions on $[0, 1]$ with finite quantile endpoints instead allow $\pi_0 = 0$ at every finite rank. Identical candidates give $B_q = \eta = \kappa = 0$. This is a sufficient coupling bound, not a claim that fitted models satisfy its population assumptions.

The condition is realizable. For $Y \sim \mathrm{Uniform}[0, 1]$, unit scale, zero raw halfwidths and centers $0, -\epsilon$, the quantiles are $Q_a(t) = t$, $Q_0(t) = t + \epsilon$. The outside indicators both equal $\mathbf{1}\{Y > t\}$, so $D = 0$. The issued intervals differ on their lower endpoint, but the score difference is the constant -2ϵ for every pool and query. Its actual variance is zero. This simple dominated-candidate example establishes nonemptiness, not useful model selection or tightness of the general bound's remaining query term.

Endpoint proximity alone does not imply a pointwise $O(\epsilon/m)$ calibration increment. Fix $\alpha = .1$, $m = 10s - 1$, $s \ge 2$, and $k = 9s$. For $Y \sim \mathrm{Uniform}[-1, 3]$, centers $0, \epsilon$ with $0 < \epsilon < 1/2$, unit scale and zero raw halfwidths, both residual quantiles are globally 4-Lipschitz. Let a retained calibration pool contain $k - 1$ zeros and $m - k$ twos. Replacing the last row from 0 to 1 changes the corrections from $(0, \epsilon)$ to $(1, 1 - \epsilon)$. At query $14/5$, both intervals miss to the right, and the paired loss changes by $4(1 - 1/\alpha)\epsilon = -36\epsilon$, independently of m. Small neighborhoods of these interior support points preserve a change of order ϵ with positive probability for each finite m. This refutes that pointwise conclusion, not an L_2 rate: the events can be rare. Better average-case coupling bounds remain possible. Exact rational checks verify both examples without refitting any historical candidate.

N.9 A NONCONSTANT TRANSPORT EXAMPLE AND ITS LIMIT

Use unit scale and a strictly increasing C^1 map ϕ such that $r_b = \phi(r_a)$ almost surely. Require ordered issued intervals for every supported correction, not merely ordered raw endpoints. Write $\psi = \phi - \mathrm{id}$, $\eta_\phi = \sup|\psi'|$, and $\kappa_\alpha = \max\{1, \alpha^{-1} - 1\}$. The score difference, in direction $b - a$, is

$$D(Z; q) = 2(w_b - w_a) + 2\psi(q) + \frac{2}{\alpha}\mathbf{1}\{r_a > q\}\{\psi(r_a) - \psi(q)\}.$$

Monotonicity gives $q_b = \phi(q)$, including ties, and identical outside indicators. The hinge is continuous; away from it, $\partial_q D = 2\psi'(q)(1 - \mathbf{1}\{r_a > q\}/\alpha)$. Thus D is $G = 2\eta_\phi\kappa_\alpha$ Lipschitz in the residual correction. With a global L_a-Lipschitz inverse CDF, finite endpoints and paired-query bound B_q, the same role-count argument gives

$$\mathrm{Var}(U_N \mid \text{fit}) \le \frac{[2B_q + GL_a\sqrt{2m/(m+1)}]^2}{2N}, \qquad N \ge m + 1. \tag{34}$$

This retains the query role and every projection order. It specializes the preceding sufficient bound, not a new concentration inequality.

Realization using only observed covariates. Let $X \sim \mathrm{Uniform}[0,1]$, $B \sim \mathrm{Bernoulli}(1/2)$ independently, and $Y = 1 + X + B$. Both centers are zero; choose

$$w_a(x) = 1, \quad w_b(x) = 1 - \epsilon \sin(2\pi x), \quad \phi(r) = r + \epsilon \sin(2\pi r), \quad 0 < \epsilon < (2\pi)^{-1}.$$

Both predictors use only X. Periodicity gives $r_a = X + B \sim \mathrm{Uniform}[0,2]$ and $r_b = \phi(r_a)$, so $L_a = 2$ and $\eta_\phi = 2\pi\epsilon$. Both issued radii are positive; the nonnegative operator is identical here because its clamp is inactive. The paired score is

$$D(Z;q) = 2\{\psi(q) - \psi(r_a)\}(1 - \mathbf{1}\{r_a > q\}/\alpha).$$

It is not constant. Its exact query envelope is $B_q = 4\epsilon\kappa_\alpha$: opposite sine extrema on the appropriate side of q attain both signs. At $\alpha = .1$ and the population quantile $q_* = 1.8$, the query derivative has zero mean but second moment $4\psi'(q_*)^2(1-\alpha)/\alpha > 0$. Calibration variation has not vanished through an invariance identity.

A finite comparison with a structure-aware ordinary bound. Put $d = m + 1$, $s = d - k$, $a_\alpha = \alpha^{-1} - 1$, and $\psi_i = \psi(r_{(i)})$. Averaging all query roles gives

$$h = \frac{2}{d}\left[k\psi_{k+1} - \sum_{i \le k} \psi_i + a_\alpha \left(\sum_{i>k} \psi_i - s\psi_k\right)\right].$$

The total absolute coefficient is $4(k + a_\alpha s)/d$. The resulting *exact* kernel envelope is $B_h = 4\epsilon(k + a_\alpha s)/d$: lower k residuals at $3/4$ and upper s at $5/4$ attain one sign; lower $1/4$ and upper $3/4$ attain the other. Their neighborhoods have positive probability. Ties cause no discrepancy because tied correction differences vanish.

Fix $\alpha = .1, m = 9999, k = 9000, N = 10^6, \delta = .05$, and $\epsilon = .01$. Then $B_h = .072$. Equation 34 and Cantelli give a one-sided radius .0120783; complete-U Hoeffding with this exact kernel range gives .0176238. Both use the same statistic, total label budget and error probability. Independently, $\pi < 22/7$ and $\sqrt{2} < 10/7$ imply the variance numerator $K < 396\epsilon$, so the squared Cantelli radius is less than $1.5\epsilon^2$. The squared range radius is $(1296/1250)\log(20)\epsilon^2 > 2\epsilon^2$. Both radii are below B_h. Exact quarter-grid interval calculations and high-precision evaluation check the algebra. The comparison is correct, but a classical ordinary bound using the same structural information is stronger.

Write $h = \sum_{j=1}^d a_j \psi(r_{(j)})$ and $\Lambda = \sum_j |a_j| = 4(k + a_\alpha s)/d$. In the complete statistic, global rank i has coefficient

$$c_i = \sum_{j=1}^d a_j \frac{\binom{i-1}{j-1}\binom{N-i}{d-j}}{\binom{N}{d}}, \qquad U_N = \sum_{i=1}^N c_i \psi(r_{(i)}), \qquad \sum_i |c_i| \le \Lambda.$$

The coefficient bound follows because each column of nonnegative subset-rank weights sums to one. It does not assume independent overlapping subsets. For $r_i = Q(V_i)$ with IID uniform V_i, suppose $f = \psi \circ Q$ is globally L_f-Lipschitz. Let $t_N = \sum_i c_i f(i/N)$ and $\Delta_N = \sup_u |F_N^V(u) - u|$. Uniform order statistics give $|U_N - t_N| \le \Lambda L_f \Delta_N$. The one-sample DKW–Massart inequality (Massart, 1990; Wei & Dudley, 2011) and its integrated tail imply

$$\Pr(\Delta_N > x) \le 2e^{-2Nx^2}, \qquad \mathbb{E}\Delta_N \le \sqrt{\pi/(2N)}.$$

The grid value t_N is not the target risk. Retaining its expectation bias gives the finite-target guarantee

$$\Pr\{|U_N - \mathbb{E}U_N| \le r_{\mathrm{DKW}}\} \ge 1 - \delta, \quad r_{\mathrm{DKW}} = \Lambda L_f \left[\sqrt{\frac{\log(2/\delta)}{2N}} + \sqrt{\frac{\pi}{2N}}\right]. \tag{35}$$

Here $\mathbb{E}U_N = R_m(b) - R_m(a)$, exactly. Neither the grid centering nor the latent uniforms need be observed to use $U_N + r_{\mathrm{DKW}}$. For this same example $\Lambda = 7.2$ and $L_f = 4\pi\epsilon$.

Bound on the same complete statistic	Radius at $\delta = .05$
Exact-range Hoeffding	.0176238
Paired Efron–Stein and Cantelli	.0120783
Structure-aware DKW	.0023628

The finite-m gain can also be bounded without numerical integration. The conditional query mean is

$$g(q) = \epsilon \left[\left(2 - \frac{2-q}{\alpha} \right) \sin(2\pi q) + \frac{\cos(2\pi q) - 1}{2\pi\alpha} \right].$$

For $Q_m = 2V$, $V \sim \mathrm{Beta}(9000, 1000)$, its mean is 1.8 and variance $.36/10001$. Also $g'(q) = (2\pi\epsilon/\alpha)(q - 1.8)\cos(2\pi q)$ and $\sup |g''| \le (2\pi\epsilon/\alpha)(1 + 3.6\pi)$ on $[0, 2]$. Taylor's theorem about the exact beta mean therefore places the benefit $\gamma = -\mathbb{E}g(Q_m)$ in $[.0108581302, .0111365420]$. Its lower endpoint exceeds $2r_{\mathrm{DKW}}$. On the event in equation 35, $U_N + r_{\mathrm{DKW}} \le -\gamma + 2r_{\mathrm{DKW}} < 0$. Thus the prespecified ordinary rule selects b with probability at least .95 under this known artificial law. This is an analytic probability over an IID assessment sample, not an executed million-row selection, fitted-model gain, or a guarantee for each separately realized final pool. A positive mean plus the paired radius does not imply that every paired-rule realization abstains. Exact subset/rank checks and separate beta integration corroborate the calculation; the Taylor enclosure supplies the power argument.

Why the condition remains restrictive. If $Y \mid X = x$ has full support on $\mathbb{R}$ for almost every x, continuity extends the residual identity to every outcome. Equal a-residuals at $z_a(x) + u$ and $z_a(x) - u$ imply $|z_a(x) + u - z_b(x)| = |z_a(x) - u - z_b(x)|$; squaring forces $z_a = z_b$. Then $\phi(r) - r = w_a(x) - w_b(x)$ on overlapping half-lines, forcing one common shift. It cancels from every calibrated radius, so issued intervals coincide. This algebraic boundary is separate from the compact-support variance assumption. The example's conditionally two-point outcomes are substantive, despite smooth marginal residuals. Finite observations cannot certify the exact transport or its global constants for an unknown fitted population.

Conditional absolute continuity alone is weaker than full support. For example, let $Y \mid X = 0$ be uniform on $[2, 3]$ and $Y \mid X = 1$ uniform on $[5, 6]$, with both input values having positive probability. Use zero centers, $w_a = 1$, $w_b(0) = 1$, and $w_b(1) = .9$. The first residual occupies $[1, 2] \cup [4, 5]$. An increasing C^1 transport equals r on the lower band and $r + .1$ on the upper, with a smooth transition in the gap. An upper-band correction changes the issued radius on lower-band queries by .1. This continuous example does not satisfy the global quantile-Lipschitz premise: its inverse CDF jumps across the gap. It corrects an overbroad impossibility claim, not the radius's applicability. Monotone transformations of a score on all hypothetical outcomes already preserve conformal prediction sets (Hanselle et al., 2025); a support-restricted residual identity need not be such a transformation of the entire prediction-set inversion.