How Synthetic Labels Improve Conformal Prediction A Perspective on Conditional Coverage

Anonymous authors
Paper under double-blind review

Abstract

Conformal prediction provides distribution-free finite-sample marginal coverage, but post-hoc calibration data may be too scarce to learn how uncertainty varies across inputs. Meanwhile, abundant covariates can often be labeled cheaply by domain models or general-purpose language models. We study whether these synthetic labels can improve conditional coverage when only a small trusted sample is available. Building on score-quantile regression, we introduce prediction-powered quantile learning: a synthetic-labeled pool estimates pinball risk, paired trusted and synthetic outcomes correct its bias, and an independent trusted split performs final conformalization. Profiling pinball risk over scalar corrections reveals that population conditional-coverage error is its functional gradient; the corresponding Hessian removes global shifts and weights remaining shape error by boundary density. Composing this geometry with prediction-powered learning yields a three-resource expansion and a benefit–cost rule for synthetic power. Across eight regression benchmarks, synthetic-powered quantile learning substantially improves downstream conditional coverage while preserving marginal validity and producing more compact prediction sets. A human-rating study finds similar gains from external LLM labels and exposes a quality–quantity–cost tradeoff.

1 Introduction

High-stakes machine learning requires reliable uncertainty estimates, not only accurate point predictions. Conformal prediction (CP) is attractive because it is distribution free, model agnostic, and gives finite-sample guarantees that a prediction set contains the future outcome at a user-specified rate Vovk et al. (2005); Shafer and Vovk (2008).

The canonical split-conformal guarantee is nevertheless marginal: it averages coverage over the test population. A procedure may therefore under-cover difficult inputs while compensating with overly wide sets on easy ones. This issue is especially important under heteroscedasticity, when predictive uncertainty varies substantially across the feature space. Although exact distribution-free feature-conditional coverage is impossible in general with finite samples Vovk (2012); Lei and Wasserman (2014); Foygel Barber et al. (2021), improving the conditional coverage behavior of conformal procedures remains a central practical objective.

Rectified Conformal Prediction (RCP) Plassier et al. (2025a) addresses this problem by learning a covariate-dependent quantile of a fixed nonconformity score, then conformalizing its residual with a scalar correction. The quantile adapts the threshold to local difficulty, while the correction restores finite-sample marginal coverage. Because pinball-loss excess risk controls score-quantile error, conditional-coverage improvement becomes a quantile-learning problem rather than a full conditional-distribution problem Plassier et al. (2025a); Chen and Li (2025).

The main obstacle is data scarcity. CP is usually applied post hoc to an already trained predictor, so the original training data may be unavailable and the only new trusted outcomes may form a modest calibration reservoir. RCP must divide this reservoir between learning a heterogeneous score threshold and applying the final conformal correction. With only tens or hundreds of trusted scores, the threshold can be unstable even when marginal coverage is reliable.

Classical semi-supervised learning offers additional covariates but no additional outcomes. In general, the marginal distribution P_X does not identify the conditional law $P_{Y|X}$: two populations may share exactly the same P_X while having arbitrarily different response mechanisms. Unlabeled covariates

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.

can therefore help only under additional structural assumptions, such as smoothness, low-density separation or cluster structure, and manifold assumptions Chapelle et al. (2006). Rather than infer $P_{Y|X}$ from P_X alone, we study a richer prediction-powered regime. We observe a small sample with trusted outcomes and a much larger sample of covariates from the same target population; an independently trained auxiliary labeler—for example, an existing domain model or a general-purpose LLM—provides synthetic outcomes at these covariates. Such labels may be biased, but they carry outcome-like information that unlabeled covariates alone do not provide.

This regime is increasingly common. Computational models can assign surrogate properties to millions of candidate molecules before expensive laboratory assays, and automated inspection systems can score sensor records before destructive verification. Recent work uses unlabeled covariates or generated scores to refine conformal sets when trusted data are scarce Seedat et al. (2023); Zhou et al. (2025); Bashari et al. (2025). Our question is different: once an independent trusted split guarantees marginal validity, can synthetic labels improve conditional coverage?

We address this question by augmenting the quantile-learning component of conformal prediction. Inspired by prediction-powered inference Angelopoulos et al. (2023a;b), we construct a bias-corrected empirical pinball risk: a large synthetic-labeled sample estimates an auxiliary risk, while paired trusted and synthetic outcomes correct its bias. We apply the idea to conditional outcome and nonconformity-score quantiles, and compare it with supervised, semi-supervised, transport, pretraining, and positive-augmentation baselines. Every method reserves an untouched trusted split for final calibration.

The final correction is not merely a validity device; it changes the value of information learned upstream. We formalize this through the profiled pinball risk obtained after optimizing over the scalar correction. Its functional gradient is the population-calibrated coverage error, and its Hessian is the density-weighted operator that removes global threshold shifts. Composing this geometry with the learner's influence function identifies the true–synthetic covariance that can affect deployment. The resulting expansion separates trusted quantile learning, synthetic-pool estimation, and final calibration, and yields an explicit power rule balancing covariance benefit against synthetic noise.

Our experiments complete this learning-to-deployment story. Across eight classical regression benchmarks and 30 common splits, synthetic-powered quantile learning substantially improves downstream conditional-coverage diagnostics and consistently produces more compact prediction sets. Intermediate pinball and post-conformal analyses connect better quantile estimation to final coverage, while a matched permutation control separates input-specific information from generic stabilization. A realistic HelpSteer2 study further shows that ratings from general-purpose LLMs can improve multivariate conditional coverage under a practical quality–quantity–cost tradeoff.

Contributions.

- We formulate a setting with abundant covariates, few trusted labels, and an auxiliary labeler, and propose prediction-powered quantile learning for conditional coverage.
- Profiling over the final scalar correction identifies calibrated coverage error as a functional gradient and the density-weighted post-conformal map as its Hessian. Combining this geometry with prediction-powered influence yields a three-resource expansion and a benefit–cost power rule.
- Across eight regression datasets and HelpSteer2, we link synthetic-powered quantile learning to deployed coverage and separate genuine pairing from generic stabilization.

1.1 Related Works

Conditional coverage and adaptive conformal prediction. Split CP guarantees marginal coverage, while exact distribution-free feature-conditional validity is generally unattainable in finite samples Vovk (2012); Lei and Wasserman (2014); Foygel Barber et al. (2021). Existing work therefore studies group-conditional guarantees, localized coverage, training-conditional guarantees, and held-out tests of conditional miscoverage Jung et al. (2023); Ding et al. (2023); Kiyani et al. (2024); Gibbs et al. (2025); Bian and Barber (2023); Hore and Barber (2025); Cauchois et al. (2021); Braun et al. (2025). Another line improves conditional behavior by changing the score or estimating a richer conditional distribution Romano et al. (2019); Lei et al. (2018); Xie et al. (2024); Izbicki et al. (2022); Plassier et al. (2025b); Colombo (2024). Most closely related, RCP estimates a conditional score

quantile and conformalizes its residual Plassier et al. (2025a). Colorful Pinball emphasizes boundary density Chen and Li (2025); we characterize how that density shapes post-conformal error.

Semi-supervised and synthetic-powered conformal prediction. Additional covariates alone do not identify $P_{Y|X}$: the same marginal P_X is compatible with arbitrarily different conditional response laws. Classical semi-supervised learning therefore relies on extra structure, such as smoothness, low-density separation or cluster assumptions, and manifold geometry Chapelle et al. (2006). Semi-supervised conformal methods use unlabeled covariates for interval refinement, reweighting, or score matching Seedat et al. (2023); Zhou et al. (2025). Synthetic-Powered Predictive Inference (SPPI) instead combines trusted and generated scores through a rank-based transporter and targets efficient calibration when trusted calibration data are scarce Bashari et al. (2025). We evaluate these approaches, but ask a different question: how can outcome-like synthetic labels improve conditional coverage when an independent trusted split is reserved for final conformalization?

Prediction-powered learning and inference. Prediction-powered inference combines a small trusted sample with abundant model predictions through a bias-corrected control variate Angelopoulos et al. (2023a;b); Ji et al. (2025). Related work studies prediction-powered optimization, semi-supervised quantile estimation, calibrated control variates, and multiple prediction sources Shoham et al. (2025). We use the same correction to learn conditional quantiles, but the distinctive challenge is downstream: only learning error that survives the final correction can improve conditional coverage.

2 Preliminaries and Problem Setting

Split conformal prediction. Let $(X, Y) \sim P$ be a future test pair. We condition on a predictive model trained independently of all samples introduced below. A fixed nonconformity score $S(x, y) \in \mathbb{R}$ measures how incompatible a candidate outcome y is with input x; for regression, a standard choice is $S(x, y) = |y - \widehat{\mu}(x)|$ or its multivariate ℓ_∞ analogue. Given an independent calibration sample $\{(X_r^c, Y_r^c)\}_{r=1}^m$, write $S_r^c = S(X_r^c, Y_r^c)$, set $\tau = 1 - \alpha$, and let $k = \lceil (m+1)\tau \rceil$. Split CP uses the kth order statistic $S_{(k)}^c$ and returns

$$\mathcal{C}_{\text{split}}(x) = \{y : S(x, y) \le S_{(k)}^c\}. \tag{1}$$

Exchangeability of the calibration and test scores gives

$$\Pr\{Y \in \mathcal{C}_{\text{split}}(X)\} \ge \frac{k}{m+1} \ge \tau, \tag{2}$$

With continuous scores, equality holds at $k/(m+1)$. We set the threshold to $+\infty$ when $k = m + 1$.

Conditional coverage and RCP. For a fitted prediction set $\mathcal{C}_\mathcal{D}$, define its conditional coverage curve by $\pi_\mathcal{D}(x) = \Pr\{Y \in \mathcal{C}_\mathcal{D}(X) \mid X = x, \mathcal{D}\}$. Marginal coverage controls only $\mathbb{E}_X \pi_\mathcal{D}(X)$. We measure conditional miscalibration through

$$\text{MSCE}_\mathcal{D} = \mathbb{E}_X[(\pi_\mathcal{D}(X) - \tau)^2], \qquad \text{HET}_\mathcal{D} = \text{Var}_X\{\pi_\mathcal{D}(X)\}. \tag{3}$$

RCP Plassier et al. (2025a) replaces the global threshold in (1) by a learned function $q(x)$. On a separate trusted calibration sample, it forms residuals $R_r^c = S_r^c - q(X_r^c)$ and lets $\widehat{\gamma}(q)$ be their kth order statistic. The resulting set is

$$\mathcal{C}_q(x) = \{y : S(x, y) \le q(x) + \widehat{\gamma}(q)\}. \tag{4}$$

Here q provides local adaptation and $\widehat{\gamma}(q)$ restores the usual marginal guarantee.

Learning the conditional score quantile. Let $F_x(s) = \Pr(S \le s \mid X = x)$ and define $q^\star(x) = F_x^{-1}(\tau)$. The oracle set $\{y : S(x, y) \le q^\star(x)\}$ has conditional coverage τ whenever F_x is continuous at $q^\star(x)$. We learn $q^\star$ through the pinball loss $\rho_\tau(u) = u\{\tau - \mathbb{1}(u < 0)\}$, with $\ell_q(x, s) = \rho_\tau\{s - q(x)\}$ and population risk $\mathcal{R}(q) = \mathbb{E}\ell_q(X, S)$.

Our analysis begins with the risk-to-coverage bridge established for RCP Plassier et al. (2025a).

Proposition 2.1 (Pinball excess risk controls pre-conformal coverage error)**.** *Assume $F_x(q^\star(x)) = \tau$ and that F_x is L_F-Lipschitz uniformly in x. Then every measurable q satisfies*

$$\mathbb{E}\big[(F_X\{q(X)\} - \tau)^2\big] \le 2L_F\{\mathcal{R}(q) - \mathcal{R}(q^\star)\}. \tag{5}$$

A proof is given in Appendix C.1. The proposition controls coverage error before the final scalar correction; Section 3.2 derives how that correction changes the relevant error.

Synthetic-supervision setting. The base predictor, auxiliary labeler, and preprocessing maps are fitted independently of the samples below and are then treated as fixed. We observe a small trusted sample $\mathcal{D}_L = \{(X_i, Y_i)\}_{i=1}^n$ and a much larger covariate-only sample $\mathcal{D}_U^X = \{\widetilde{X}_j\}_{j=1}^N$, where $N \gg n$. The covariates in both samples are independent draws from the same target distribution P_X, but true outcomes are observed only in $\mathcal{D}_L$.

The auxiliary labeler supplies $Y_i^\dagger$ at each trusted covariate and $\widetilde{Y}_j$ at each unlabeled covariate. We write $S_i = S(X_i, Y_i)$ for the trusted score, $S_i^\dagger = S(X_i, Y_i^\dagger)$ for the paired synthetic score, and $\widetilde{S}_j = S(\widetilde{X}_j, \widetilde{Y}_j)$ for the synthetic-pool score. The same labeler is applied in both places, so $(X, S^\dagger) \overset{d}{=} (\widetilde{X}, \widetilde{S})$, although this common synthetic law need not equal the trusted law of (X, S). Finally, an independent trusted sample $\mathcal{D}_C = \{(X_r^c, Y_r^c)\}_{r=1}^m$ is reserved exclusively for the scalar correction in (4). Thus trusted learning data identify the target risk, the large synthetic pool estimates an auxiliary risk, and untouched trusted calibration data restore finite-sample coverage.

3 Methodology and Theory

Roadmap. We proceed in three steps. First, we define prediction-powered quantile learning and show that an untouched calibration split preserves exact marginal validity. Second, we profile pinball risk over the final scalar correction, revealing which threshold errors remain visible after conformalization. Third, we pass prediction-powered learning fluctuations through this geometry to obtain the three-resource expansion and its synthetic-power rule. Appendix B states the regularity and secondary results, while Appendix C.4 supplies the learner construction.

3.1 Prediction-powered quantile learning

Let P_n denote the empirical average over $\mathcal{D}_L$ and $\widetilde{P}_N$ the empirical average over the synthetic pool. We fit the conditional score quantile with

$$\widehat{\mathcal{R}}_\lambda(q) = P_n \ell_q(X, S) + \lambda\{\widetilde{P}_N \ell_q(\widetilde{X}, \widetilde{S}) - P_n \ell_q(X, S^\dagger)\}. \tag{6}$$

The choice $\lambda = 0$ recovers supervised learning, while $\lambda = 1$ gives unit PPI. Because paired and pool synthetic scores share a law, $\mathbb{E}\widehat{\mathcal{R}}_\lambda(q) = \mathcal{R}(q)$ for every q and λ, even under a biased labeler. The subtraction therefore removes the population bias of the pool term.

For comparison, positive augmentation omits the paired subtraction and may bias the population threshold, although untouched calibration still preserves marginal validity for the final prediction set.

Independent final calibration separates marginal validity from the quality of the learned threshold. We first formalize this protection, then decompose the conditional error that remains.

Proposition 3.1 (Exact validity under adaptive synthetic learning)**.** *Let $\widehat{q}$ and any power or tuning rule be measurable functions of $(\mathcal{D}_L, \mathcal{D}_U^X)$ and randomness independent of $\mathcal{D}_C$ and the test pair. If* (4) *uses the exact kth calibration order statistic, then*

$$\Pr\{Y \in \mathcal{C}_{\widehat{q}}(X) \mid \mathcal{D}_L, \mathcal{D}_U^X\} \geq k/(m+1) \geq \tau. \tag{7}$$

Under continuous residuals, equality holds at $k/(m+1)$.

Thus synthetic learning and adaptive power selection may change conditional efficiency without invalidating marginal coverage, provided the final calibration sample remains untouched. Validity alone, however, does not describe a fitted procedure's conditional behavior.

Proposition 3.2 (Realized conditional-error decomposition)**.** *For a fitted procedure, the mean coverage $U_\mathcal{D} = \mathbb{E}_X \pi_\mathcal{D}(X)$ satisfies*

$$\mathrm{MSCE}_\mathcal{D} = \mathrm{HET}_\mathcal{D} + (U_\mathcal{D} - \tau)^2. \tag{8}$$

If the marginal residual CDF is continuous, then conditional on $\widehat{q}$,

$$U_\mathcal{D} \overset{d}{=} U_{(k)} \sim \mathrm{Beta}(k, m+1-k), \qquad \mathbb{E}[(U_{(k)} - \tau)^2] =: \sigma_{m,\tau}^2. \tag{9}$$

Averaging over fitted procedures also gives

$$\mathbb{E}_{\mathcal{D}}\,\mathrm{MSCE}_{\mathcal{D}} = \mathbb{E}_X[(\mathbb{E}_{\mathcal{D}}\pi_{\mathcal{D}}(X) - \tau)^2] + \mathbb{E}_X\,\mathrm{Var}_{\mathcal{D}}\{\pi_{\mathcal{D}}(X)\}. \tag{10}$$

Together, the identities separate covariate heterogeneity from finite-calibration offset and require evaluation within each fitted run before aggregation across repetitions.

3.2 What survives final conformalization?

Final conformalization treats threshold level and shape differently: it cancels every constant shift exactly at finite sample, so only covariate-dependent shape can alter the deployed set.

Proposition 3.3 (Global shift invariance)**.** *For every measurable q and constant c,*

$$\widehat{\gamma}(q + c) = \widehat{\gamma}(q) - c, \qquad \mathcal{C}_{q+c}(x) = \mathcal{C}_q(x). \tag{11}$$

Thus deployment depends on the shape of q, not its global level. We now make this quotient structure explicit at the population level. Define the profiled correction and risk

$$\gamma_q \in \operatorname*{arg\,min}_{\gamma\in\mathbb{R}} \mathcal{R}(q + \gamma), \qquad \overline{\mathcal{R}}(q) = \mathcal{R}(q + \gamma_q), \qquad \pi_q(x) = F_x\{q(x) + \gamma_q\}. \tag{12}$$

When the residual CDF is continuous, the first-order condition gives $\mathbb{E}_X \pi_q(X) = \tau$.

Proposition 3.4 (Profiled risk and post-conformal coverage)**.** *Assume that γ_q is locally unique and differentiation may pass through expectation. For every constant c and admissible direction h,*

$$\overline{\mathcal{R}}(q + c) = \overline{\mathcal{R}}(q), \qquad D\overline{\mathcal{R}}(q)[h] = \mathbb{E}_X[\{\pi_q(X) - \tau\}h(X)], \tag{13}$$

$$\|\nabla\overline{\mathcal{R}}(q)\|^2_{L_2(P_X)} = \mathbb{E}_X[\{\pi_q(X) - \tau\}^2]. \tag{14}$$

If $f_q(x) = f_{S|X}\{q(x) + \gamma_q \mid x\}$ exists and $\bar{f}_q = \mathbb{E}f_q(X) > 0$, then

$$(\mathcal{H}_q h)(x) = f_q(x)\left[h(x) - \frac{\mathbb{E}\{f_q(X)h(X)\}}{\bar{f}_q}\right], \qquad \mathcal{H}_q = \nabla^2\overline{\mathcal{R}}(q). \tag{15}$$

Thus population-calibrated conditional error is the functional gradient of a shift-invariant risk, and the density-weighted post-conformal operator is its curvature. Appendix B.1 records the derivative of γ_q, positive semidefiniteness, and the quotient-space quadratic form. Applying Proposition 2.1 to $q + \gamma_q$ gives

$$\|\nabla\overline{\mathcal{R}}(q)\|_2^2 \le 2L_F\{\overline{\mathcal{R}}(q) - \mathcal{R}(q^\star)\}. \tag{16}$$

The inequality controls magnitudes, not rankings: profiled risk and conditional-coverage error can order the same candidates differently.

For a target level u, let $\Gamma(q, u)$ solve $\mathbb{E}_X F_X\{q(X) + \Gamma(q, u)\} = u$ and set $\mathcal{P}(q, u)(x) = F_x\{q(x) + \Gamma(q, u)\}$. At $q^\star$, the profiled Hessian gives the local deployment map.

Corollary 3.5 (Local density-weighted projection)**.** *Set $f_\star(x) = f_{S|X}\{q^\star(x) \mid x\}$ and $\bar{f}_\star = \mathbb{E}f_\star(X) > 0$. Under the smoothness conditions of Appendix B.1,*

$$\mathcal{P}(q^\star + e, \tau + v) - \tau = \mathcal{L}e + \frac{f_\star}{\bar{f}_\star}v + r_{e,v}, \qquad \|r_{e,v}\|_2 \le C(\|e\|_\infty + |v|)^2, \tag{17}$$

$$(\mathcal{L}e)(x) = f_\star(x)\left[e(x) - \frac{\mathbb{E}\{f_\star(X)e(X)\}}{\bar{f}_\star}\right]. \tag{18}$$

The density converts threshold error into coverage error, while the scalar correction removes its density-weighted global component. Thus $\mathcal{L} = \mathcal{H}_{q^\star}$ acts on threshold shape modulo constant shifts.

To connect this function-space derivative to a learned model, let $q_0 = q_{\theta_0}$ be an interior stationary point of the trusted pinball risk, and let γ_0 satisfy $\mathbb{E}_X F_X\{q_0(X) + \gamma_0\} = \tau$. Write $\pi_0(x) = F_x\{q_0(x) + \gamma_0\}$, $d(x) = \nabla_\theta q_\theta(x)|_{\theta_0}$, $J = \nabla^2_\theta \mathcal{R}(q_\theta)|_{\theta_0}$, and

$$f_c(x) = f_{S|X}\{q_0(x) + \gamma_0 \mid x\}, \qquad d_c(x) = d(x) - \frac{\mathbb{E}\{f_c(X)d(X)\}}{\mathbb{E}f_c(X)}. \tag{19}$$

When the tangent contains an unpenalized intercept, stationarity already implies $\gamma_0 = 0$; we retain it to cover restricted or penalized learners.

Definition 3.6 (Post-conformal influence map)**.** For a learning-score perturbation v, define

$$(\mathsf{B}v)(x) = f_c(x)d_c(x)^\top J^{-1}v, \qquad G_c = J^{-1}\mathbb{E}[f_c(X)^2 d_c(X)d_c(X)^\top]J^{-1}. \tag{20}$$

Then $\|\mathsf{B}v\|^2_{L_2(P_X)} = v^\top G_c v$.

The inverse Hessian maps a score perturbation to a threshold change, while $f_c d_c$ retains only the component visible after conformalization; their composition $\mathsf{B}v$ is the deployed coverage perturbation.

3.3 When do synthetic labels help?

At the learning boundary define

$$\xi = \{\mathbb{1}(S \le q_0(X)) - \tau\}d(X), \qquad \xi^\dagger = \{\mathbb{1}(S^\dagger \le q_0(X)) - \tau\}d(X), \tag{21}$$

with pool analogue $\widetilde{\xi}$. The deployment-scale constants are

$$a = \mathrm{tr}\{G_c \,\mathrm{Var}(\xi)\}, \qquad b = \mathrm{tr}\{G_c \,\mathrm{Cov}(\xi, \xi^\dagger)\}, \qquad c = \mathrm{tr}\{G_c \,\mathrm{Var}(\xi^\dagger)\}, \tag{22}$$

and $C_{\mathrm{cal}} = \mathbb{E}[(f_c(X)/\mathbb{E}f_c(X))^2]$. Thus a, b, and c respectively quantify trusted deployment variance, usable true–synthetic covariance, and the synthetic variance cost.

Appendix B, Assumption B.1, states the fixed-dimensional conditions: stationary curvature, a uniform asymptotic-linear learner, a smooth $L_2(P_X)$ deployment map, independent samples, and proportional rates. Appendix C.4 constructs a one-step estimator.

Define

$$L(\lambda) = \frac{a - 2b\lambda + (1+\kappa)c\lambda^2}{n}, \qquad \sigma^2_{m,\tau} = \mathbb{E}[(U_{(k)} - \tau)^2]. \tag{23}$$

Theorem 3.7 (Three-resource post-conformal expansion)**.** *Under Assumption B.1, the following expansions hold uniformly over the fixed admissible power set:*

$$\mathbb{E}_{\mathcal{D}}\|\pi_{\mathcal{D},\lambda} - \pi_0\|_2^2 = L(\lambda) + C_{\mathrm{cal}}\sigma^2_{m,\tau} + o(n^{-1} + N^{-1} + m^{-1}), \tag{24}$$

$$\mathcal{V}_{\mathrm{proc}}(\lambda) = L(\lambda) + C_{\mathrm{cal}}\,\mathrm{Var}(U_{(k)}) + o(n^{-1} + N^{-1} + m^{-1}). \tag{25}$$

The expansion separates three resources. Trusted learning contributes a/n; paired covariance supplies the benefit $-2b\lambda/n$; paired and pool estimation incur $(1+\kappa)c\lambda^2/n$. Final calibration contributes a separate order-$1/m$ term that no synthetic pool can remove.

The covariance benefit has two distinct sources. Appendix B, Proposition B.2, decomposes it as

$$b = b_{\mathrm{res}} + b_{\mathrm{align}}, \tag{26}$$

where b_{res} is within-X covariance of true and synthetic boundary crossings, while b_{align} is between-X alignment of their systematic score errors. Exact specification removes alignment; conditional independence removes residual coupling. When both hold, Corollary B.3 gives $b = 0$ and zero optimal nonnegative PPI power. This limits the matched bias-robust correction; it does not preclude direct pooling under a certified synthetic law.

Corollary 3.8 (Benefit–cost rule for synthetic power)**.** *For $c > 0$ and admissible interval Λ,*

$$\lambda^\star = \Pi_\Lambda\left\{\frac{b}{(1+\kappa)c}\right\}. \tag{27}$$

When clipping is inactive,

$$L(\lambda) - L(\lambda^\star) = \frac{(1+\kappa)c}{n}(\lambda - \lambda^\star)^2. \tag{28}$$

Unit power improves on supervision if and only if $2b > (1+\kappa)c$.

The rule gives a direct interpretation: b is the usable covariance benefit, c is the synthetic-variance price, and $\kappa = n/N$ records pool-estimation precision. Positive projected covariance supports some positive power but does not justify unit power. Appendix B gives exact-specification (Corollary B.4), robustness (Proposition B.7), estimable-power (Corollary B.8), and allocation (Corollary B.9) results.

4 EXPERIMENTS

We now ask whether the proposed use of synthetic labels improves conditional coverage in practice. The main benchmark compares the complete baseline set on eight regression datasets and 30 common data splits. All methods use the same data roles, and every synthetic method is compared with the supervised method in the same conformal family. Additional experiments isolate genuine synthetic-label information and study labeler quality, synthetic-pool size, and trusted-label scarcity.

4.1 EXPERIMENTAL SETUP

Data and label budget. We evaluate eight regression benchmarks: Bike Sharing (Bike), Diamonds (Diam.), Gas Turbine (Gas), Naval Propulsion (Naval), SGEMM, Superconductivity (Cond.), Transcoding (Trans.), and WEC. They range from scalar outcomes to the 49-dimensional WEC response; all cleaned observations are used except for a fixed 50,000-row SGEMM sample. In the main low-label regime, 20% of the observations train and validate the auxiliary labeler, 30% supply covariates for the synthetic pool, 2% train the base predictor or CQR endpoints, and only 1% form the conformal reservoir. RCP divides this reservoir equally between score-quantile learning and untouched final calibration. Independent samples fit conditional-coverage groups, select worst slices, and evaluate every method, while a separate preprocessing role fits imputation and scaling. Appendix A.1 gives dimensions, sources, and complete role accounting.

Models and methods. All CQR and RCP variants use the same two-hidden-layer width-128 ReLU architecture, common initialization within each dataset–seed family, and the same optimization budget; only the learning objective and the synthetic data available to it change. The auxiliary labeler is an externally trained random forest, and all methods use exact order-statistic calibration.

Split is ordinary split conformal prediction using trusted residual scores. SPPI Bashari et al. (2025) transports ranks between trusted and synthetic score distributions, whereas NNM Zhou et al. (2025) corrects synthetic scores using nearby trusted observations before calibration. CQR predicts lower and upper conditional outcome quantiles and conformalizes their endpoint error. CQR-PPI keeps the same architecture and conformalization step but replaces supervised endpoint pinball risk by (6).

RCP learns the conditional τ-quantile of the residual score and then applies an independent scalar correction. RCP-PTFT first trains this quantile network on synthetic scores and then fine-tunes it on trusted scores. RCP-PPI-CV fits prediction-powered candidates and selects $\lambda \in \{0, .25, .5, .75, 1\}$ using trusted pinball validation. RCP-Aug instead minimizes the positively weighted objective $P_n \ell_q(X, S) + 0.5\widetilde{P}_N \ell_q(\widetilde{X}, \widetilde{S})$. It is a useful empirical comparator, but unlike PPI it has no paired subtraction and can therefore change the population threshold when the synthetic law is biased. Our main RCP method, RCP-PPI, optimizes the bias-corrected objective (6) at $\lambda = 1$. The primary comparisons—CQR versus CQR-PPI and RCP versus RCP-PPI—hold architecture, initialization, data roles, optimization, and final calibration fixed, changing only the quantile-learning objective.

Metrics and comparisons. We evaluate grouped mean-squared conditional error (MSCE), Worst-Slice Coverage (WSC), and L_1 Excess Risk of Target coverage (L_1-ERT). Independent data define $K = 30$ covariate groups, and held-out observations estimate $\sum_k \widehat{p}_k(\widehat{c}_k - \tau)^2$. Lower MSCE and L_1-ERT and higher WSC are better. The main text reports complete MSCE and WSC tables; Appendix A.4 reports L_1-ERT, paired effects, marginal coverage, log volume, and a split-half MSCE audit. Entries are mean $\pm$ standard deviation over 30 common splits, and paired intervals always compare methods within the same dataset and conformal family.

4.2 MAIN RESULTS

Tables 1 and 2 report the complete ten-method benchmark under the moderate synthetic labeler; Appendix A.4 gives the corresponding L_1-ERT table and paired uncertainty.

Prediction-powered quantile learning improves downstream conditional coverage. Relative to CQR, CQR-PPI lowers grouped MSCE and raises WSC on seven of eight datasets; the complete L_1-ERT table in Appendix A.4 shows the same seven-dataset pattern. The strongest MSCE changes include $3.89 \to 1.04$ on SGEMM, $7.47 \to 5.54$ on Naval, and $4.12 \to 2.59$ on Transcoding. Relative to RCP,

Table 1: Grouped mean-squared conditional-coverage error (MSCE; lower is better) on Bike Sharing, Diamonds, Gas Turbine, Naval Propulsion, SGEMM, Superconductivity, Transcoding, and WEC. We report $10^3\times$ mean $\pm$ standard deviation over 30 common splits. Rules separate method families; bold marks the best value in each column, and abbreviated headers are defined in Section 4.1.

Method	Bike	Diam.	Gas	Naval	SGEMM	Cond.	Trans.	WEC
Split	5.89±1.58	13.10±2.62	8.13±2.08	10.15±3.87	5.94±1.53	8.12±2.55	5.68±1.83	13.28±4.36
SPPI	6.92±2.18	14.01±2.60	10.05±2.54	11.99±4.83	6.16±1.23	11.05±4.32	6.38±1.71	15.98±4.79
NNM	6.13±2.09	12.99±2.85	8.16±2.00	11.02±4.86	5.86±1.04	9.02±3.67	5.64±1.26	13.13±4.35
CQR	3.39±1.19	6.06±1.66	2.45±0.71	7.47±2.84	3.89±1.03	6.07±2.00	4.12±1.44	8.85±2.73
CQR-PPI	2.60±1.68	3.36±0.74	**2.14±0.62**	**5.54±1.98**	**1.04±0.25**	4.46±1.28	**2.59±0.58**	9.86±3.34
RCP	5.76±3.16	7.12±2.31	4.42±2.19	10.29±5.74	3.68±1.61	7.02±2.55	4.75±2.36	6.71±3.44
RCP-PTFT	4.98±2.64	5.58±1.88	3.43±1.48	9.27±5.33	3.17±1.54	6.48±2.15	4.19±2.46	**5.38±2.54**
RCP-PPI-CV	2.90±1.56	4.35±2.43	3.64±1.57	7.64±4.66	2.78±2.20	4.29±1.86	3.78±2.98	7.88±3.58
RCP-Aug	3.40±1.62	3.67±0.88	3.21±1.38	7.60±3.86	2.26±1.09	4.38±1.04	3.50±2.46	5.92±2.98
RCP-PPI	**2.52±1.55**	**3.33±1.22**	3.01±1.17	6.41±2.57	3.09±2.13	**4.07±1.98**	3.20±2.81	8.00±3.39

Table 2: Worst-Slice Coverage (WSC) over the same 30 common splits; higher is better. Rules separate method families, and bold marks the best value in each dataset.

Method	Bike	Diam.	Gas	Naval	SGEMM	Cond.	Trans.	WEC
Split	.843±.048	.620±.051	.719±.057	.752±.080	.739±.049	.790±.043	.745±.050	.781±.040
SPPI	.817±.053	.604±.048	.686±.059	.705±.080	.737±.052	.758±.055	.732±.049	.758±.045
NNM	.838±.050	.621±.052	.723±.061	.731±.088	.741±.056	.785±.051	.746±.047	.785±.047
CQR	.867±.047	.763±.046	.827±.035	.817±.062	.784±.036	.818±.052	.779±.049	.821±.035
CQR-PPI	.883±.044	.810±.028	**.854±.032**	**.828±.057**	**.875±.021**	.847±.043	.824±.034	.810±.042
RCP	.855±.072	.724±.056	.803±.066	.760±.098	.797±.038	.802±.066	.795±.039	.808±.052
RCP-PTFT	.857±.073	.761±.052	.827±.046	.779±.089	.817±.035	.810±.057	.809±.035	**.822±.046**
RCP-PPI-CV	**.891±.053**	.798±.056	.821±.059	.817±.083	.829±.049	.843±.046	.813±.048	.813±.053
RCP-Aug	.882±.059	.803±.031	.839±.048	.807±.080	.835±.034	.846±.051	.827±.036	.818±.053
RCP-PPI	.890±.051	**.816±.041**	.839±.049	.825±.068	.837±.041	**.851±.044**	**.832±.048**	.813±.048

RCP-PPI lowers grouped MSCE on seven datasets, raises WSC on all eight, and lowers L_1-ERT on all eight. On Diamonds, for example, MSCE falls from 7.12 to 3.33 while WSC rises from .724 to .816. Its validation pinball risk is also lower on every dataset (Appendix A.7), supporting the intended chain from improved quantile learning to post-conformal conditional coverage. WEC is the principal MSCE exception: RCP-PTFT is strongest there, although both PPI variants remain competitive under WSC and L_1-ERT.

The insertion point and robustness correction matter. SPPI and NNM enlarge the calibration resource but do not explicitly learn heterogeneous score quantiles; under this protocol their conditional metrics are generally close to, or worse than, Split CP. Synthetic pretraining and positive augmentation frequently improve RCP, but neither is bias robust. RCP-PPI-CV also improves many cells, yet it is less consistent than fixed unit PPI, matching the concern that power selection can be noisy when trusted validation data are scarce. These comparisons motivate controlled CQR/CQR-PPI and RCP/RCP-PPI contrasts that differ only in the quantile-learning objective.

A matched permutation experiment further separates input-specific synthetic information from generic regularization. On five prespecified datasets, genuine augmentation lowers both the raw and split-half grouped MSCE and raises WSC relative to a placebo that preserves the synthetic-score marginal, sample size, network, and optimizer while breaking covariate–score pairing (Appendix A.5). The experiment concerns positive augmentation rather than the bias-corrected PPI objective, but it shows that synthetic supervision can contribute conditional information beyond threshold stabilization.

The improvements are not purchased by indiscriminately wider sets. Across all method–dataset cells, marginal coverage remains in $[0.882, 0.919]$ and no full prediction set occurs. CQR-PPI and RCP-PPI reduce mean log volume on every dataset relative to their supervised counterparts, while RCP-PPI also improves every WSC and L_1-ERT point estimate. Appendix A.4 reports the full coverage, volume, empty-set, paired-effect, and split-half audits. Resource sweeps further show that gains are largest when trusted labels are scarce and tend to plateau once the synthetic-pool estimation term is small, matching the three-resource expansion. Teacher quality is not reducible to global predictive

accuracy: stronger teachers help some datasets but not others, consistent with the theory's emphasis on boundary-aligned post-conformal information rather than generic label fidelity (Appendix A.6).

We finally replace the tabular teacher with external language models. HelpSteer2 supplies five human ratings–helpfulness, correctness, coherence, complexity, and verbosity–for each prompt–response pair. Over 20 paired seeds, equal-count Flash-PPI and Pro-PPI reduce grouped MSCE relative to supervised RCP from 0.00665 to 0.00502 and 0.00400, improvements of 24.6% and 39.9%. Pro-PPI also raises minimum-group coverage from 0.738 to 0.773 and reduces mean log volume from -0.426 to -0.481. Equal-count Pro costs 3.20 times as much as Flash; at nearly equal spend, 400 Flash labels yield slightly lower average MSCE and smaller sets than 54.8 Pro labels, whereas Pro retains slightly higher minimum-group coverage. Figure 1 summarizes this quality–quantity–cost tradeoff, and Appendix A.8 reports the prompt-level split, boundary audit, recorded costs, and paired uncertainty.

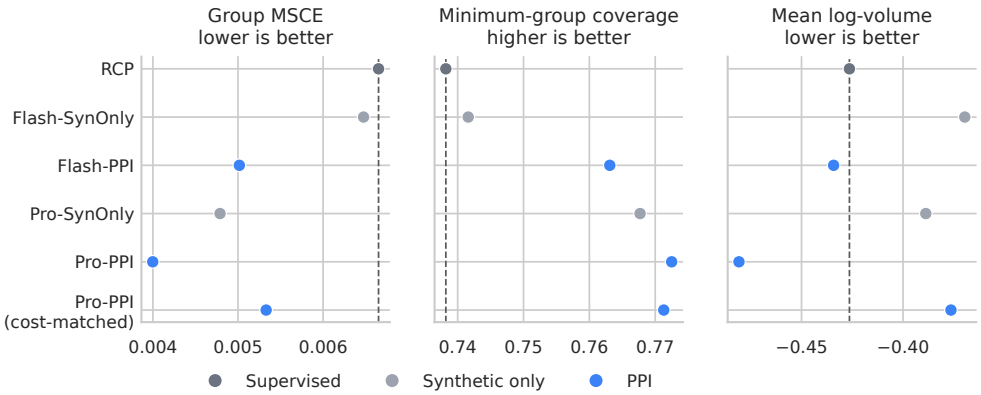

Figure 1: HelpSteer2 downstream results over 20 paired seeds. Equal-count PPI uses 400 auxiliary labels; the cost-matched Pro arm uses 54.8 on average. Dashed lines mark supervised RCP.

Two further diagnostics clarify the profiled-risk theory without changing the method. Across 240 candidate–shift audits, global shifts changed raw-pinball selection 205 times and the complete raw-risk ordering 235 times, whereas cross-calibrated criteria and deployed thresholds were invariant up to 8.9×10^{-16}. A separate four-dataset study found that coverage-moment selection did not consistently improve over fixed unit PPI, so we retain the pre-specified method rather than selecting a favorable rule post hoc. Appendix A.11 reports this negative result together with the finite-sample unbiasedness and variance analysis of the quadratic coverage-discrepancy estimator.

Agreement across metrics is informative: MSCE averages group deviations, WSC targets the worst independently selected region, and L_1-ERT measures predictable coverage heterogeneity. RCP-PPI improves every WSC and L_1-ERT point estimate, including on WEC, where grouped MSCE is the main exception, so the signal is not tied to one clustering summary. RCP-Aug sometimes attains lower MSCE but lacks PPI's population bias cancellation, highlighting a tradeoff between flexible augmentation and bias robustness rather than universal dominance.

5 Conclusion

With abundant covariates but few trusted labels, prediction-powered quantile learning offers a principled way to turn synthetic supervision into more reliable conditional coverage while preserving the usual marginal guarantee. Across external labelers—including existing domain models and general-purpose LLMs—our experiments show that bias-corrected quantile learning improves deployed coverage and often yields smaller prediction sets, providing a new post-hoc route for integrating synthetic labels into conformal prediction. The theory explains why these gains are not determined by generic label accuracy or raw learning-stage covariance: final conformalization removes global threshold shifts and retains only density-weighted shape information aligned with trusted learning error that remains visible at deployment. This post-conformal perspective separates usable synthetic information from synthetic noise and finite-calibration error, and supplies an explicit benefit–cost rule for how strongly synthetic labels should influence quantile learning.

AI USE STATEMENT

Large language models were used in two distinct roles. First, designated external LLMs generated synthetic labels for the HelpSteer2 experiment; these outputs were treated as experimental inputs and evaluated under the protocol described in Section 4 and Appendix A.8. Second, generative AI tools assisted with checking mathematical derivations and proofs and with polishing the manuscript for clarity, grammar, and consistency. The authors independently verified every theorem statement, proof, experimental procedure, numerical result, and scientific conclusion and take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

Complete assumptions and proofs are provided in the appendices. The supplementary package contains experimental configurations, data-role definitions, outcome-access audits, raw task outputs, aggregation and plotting scripts, and the additional profiled-risk diagnostics. Exact order-statistic calibration, random seeds, model architectures, optimization settings, and metric definitions are documented in the appendix and configuration files.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a. doi: 10.1126/science.adi6000.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

Meshi Bashari, Roy Maor Lotan, Yonghoon Lee, Edgar Dobriban, and Yaniv Romano. Synthetic-powered predictive inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3925.

Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023.

Sacha Braun, David Holzmüller, Michael I. Jordan, and Francis Bach. Conditional coverage diagnostics for conformal prediction, 2025.

Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22 (81):1–42, 2021.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

Qianyi Chen and Bo Li. Colorful pinball: Density-weighted quantile regression for conditional guarantee of conformal prediction. *arXiv preprint arXiv:2512.24139*, 2025.

Nicolò Colombo. Normalizing flows for conformal regression. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 881–893. PMLR, 2024.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4): 1100–1126, 2025. doi: 10.1093/jrsssb/qkaf008.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: Randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):549–578, 2025. doi: 10.1093/jrsssb/qkae103.

Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of ai. *arXiv preprint arXiv:2501.09731*, 2025.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023.

Shayan Kiyani, George J. Pappas, and Hamed Hassani. Conformal prediction with learned features. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 24749–24769. PMLR, 2024.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

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.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 49459–49492. PMLR, 2025a.

Vincent Plassier, Alexander Fishkov, Mohsen Guizani, Maxim Panov, and Eric Moulines. Probabilistic conformal prediction with approximate conditional validity. In *International Conference on Learning Representations*, 2025b.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Nabeel Seedat, Alan Jeffares, Fergus Imrie, and Mihaela van der Schaar. Improving adaptive conformal prediction using self-supervised learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10160–10177. PMLR, 2023.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

Noa Shoham, Ron Dorfman, Shalev Shaer, Kfir Yehuda Levy, and Yaniv Romano. Prediction-powered semi-supervised learning with online power tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3632.

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*, pages 475–490. PMLR, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Ran Xie, Rina Barber, and Emmanuel Candes. Boosted conformal prediction intervals. *Advances in Neural Information Processing Systems*, 37:71868–71899, 2024.

Xuanning Zhou, Hao Zeng, Xiaobo Xia, Bingyi Jing, and Hongxin Wei. Semi-supervised conformal prediction with unlabeled nonconformity score. *arXiv preprint arXiv:2505.21147*, 2025.

A Experimental Details and Full Results

A.1 Datasets and role protocol

The eight datasets are Bike Sharing, Diamonds, Gas Turbine, Naval Propulsion, SGEMM, Superconductivity, Transcoding, and WEC. They span scalar and multivariate regression, with response dimension up to 49. The package records the source, version, preprocessing, and license of every dataset. All cleaned observations are used except for a fixed 50,000-row SGEMM sample.

Each seed begins with one random permutation and non-overlapping maximum reservoirs: 1% preprocessing, 16% auxiliary-labeler training, 4% labeler validation, up to 40% synthetic-pool covariates, up to 4% base/CQR training, up to 2% conformal data, 3% grouping, 3% worst-slice selection, and the remaining observations for evaluation. The main experiment uses a 30% synthetic pool, 2% base training, and a 1% conformal reservoir. Resource sweeps use nested prefixes, so changing one resource does not move the evaluation data or any other role. Ordinary synthetic methods never access true outcomes in the synthetic pool.

A.2 Detailed baseline definitions

Split, SPPI, and NNM. Split CP calibrates the global ℓ_∞ residual score on the complete conformal reservoir. SPPI Bashari et al. (2025) uses the published rank-based transporter between trusted and synthetic score distributions. NNM Zhou et al. (2025) matches synthetic observations to trusted neighbors and corrects their scores before calibration. All three share the same base predictor, data roles, and evaluation rows.

CQR and CQR-PPI. CQR Romano et al. (2019) predicts coordinatewise lower and upper conditional outcome quantiles with a two-hidden-layer network and conformalizes the maximum signed endpoint error. CQR-PPI replaces each endpoint pinball risk by (6), using paired and pool synthetic outcomes; both methods run on all eight datasets, WEC included.

RCP family. RCP learns the conditional τ-quantile of the ℓ_∞ residual score on one half of the conformal reservoir and uses the other half for final calibration. RCP-PTFT pretrains the same network on synthetic scores and fine-tunes on trusted scores. RCP-PPI-CV chooses λ from $\{0, .25, .5, .75, 1\}$ by five-fold trusted pinball validation and refits on the full threshold-learning half. RCP-Aug minimizes $P_n\ell_q(X, S) + 0.5\widetilde{P}_N\ell_q(\widetilde{X}, \widetilde{S})$; because it has no paired subtraction, it is not bias robust. RCP-PPI trains the same network with (6) at $\lambda = 1$. All RCP variants use bounded nonnegative outputs, Adam learning rate 2×10^{-3}, 100 epochs, weight decay, gradient clipping, and common initialization. CQR uses the same width-128 architecture for 200 epochs.

Ablation controls. RCP-Aug-Permuted permutes synthetic scores across pool covariates. RCP-Extra uses true pool outcomes and is explicitly a resource upper control. RCP-Pool assigns equal positive weight to every trusted and synthetic score and represents stronger trust in the synthetic law. None of these controls appears in the ten-method ranking.

A.3 Metrics and statistical reporting

K-means with $K = 30$ is fitted only on the independent grouping sample. For held-out group coverage $\widehat{c}_k$ and group mass $\widehat{p}_k$, the main table reports

$$\widehat{\mathrm{MSCE}}_{30} = \sum_{k=1}^{30} \widehat{p}_k(\widehat{c}_k - \tau)^2. \tag{29}$$

This nonnegative estimator contains finite-evaluation Bernoulli noise. We therefore repeat the analysis with $K = 10$, split the evaluation rows in each group into independent halves A_k, B_k, and report

$$\widehat{\mathrm{MSCE}}_{\mathrm{split}} = \sum_{k=1}^{10} \widehat{p}_k(\widehat{c}_{A_k} - \tau)(\widehat{c}_{B_k} - \tau), \tag{30}$$

whose expectation removes the leading Bernoulli term conditional on the groups. WSC searches 64 fixed projection directions on a dedicated selection sample, subject to minimum mass 0.1, and

Table 3: L_1 Excess Risk of Target coverage (L_1-ERT) over 30 common splits. Entries are mean $\pm$ standard deviation; lower is better.

Method	Bike	Diam.	Gas	Naval	SGEMM	Cond.	Trans.	WEC
Split	.064±.008	.127±.006	.052±.005	.062±.017	.112±.005	.095±.013	.073±.004	.112±.010
SPPI	.066±.009	.129±.004	.058±.007	.063±.017	.111±.005	.108±.017	.075±.005	.120±.010
NNM	.062±.008	.127±.005	.053±.006	.064±.016	.112±.005	.098±.017	.073±.005	.113±.010
CQR	.039±.009	.085±.010	.031±.008	.047±.012	.087±.009	.070±.011	.070±.007	.090±.008
CQR-PPI	.035±.014	.062±.006	**.028±.009**	**.042±.012**	.034±.007	**.056±.009**	.059±.004	.092±.008
RCP	.058±.014	.090±.008	.044±.010	.065±.021	.065±.010	.081±.010	.053±.009	.091±.010
RCP-PTFT	.052±.012	.075±.009	.039±.009	.059±.023	.051±.011	.075±.008	.049±.007	.081±.008
RCP-PPI-CV	.038±.016	.063±.019	.039±.011	.054±.020	.034±.015	.063±.011	.040±.011	.077±.010
RCP-Aug	.041±.013	.059±.008	.037±.009	.051±.017	.039±.011	.066±.007	.042±.008	**.074±.009**
RCP-PPI	**.033±.017**	**.051±.009**	.036±.009	.045±.015	**.031±.010**	.060±.008	**.034±.007**	.078±.008

evaluates the selected slab on different observations. L_1-ERT uses an unweighted cross-fitted logistic coverage model against the constant nominal-coverage predictor. We also report marginal coverage, original-scale log volume, empty/full-set frequencies, failures, and true-outcome access.

Every main result uses 30 common data splits. Paired effects subtract the supervised method in the same family on the same split and evaluation observations. Percentile intervals use 5,000 bootstrap resamples of the within-split difference. They are descriptive, dataset-specific intervals without multiple-comparison adjustment; no cross-dataset effect is formed.

A.4 Paired results, marginal coverage, and volume

The main text reports complete MSCE and WSC results; this section adds L_1-ERT, paired grouped-MSCE changes, marginal coverage, mean log volume, and an independent split-half audit.

Table 4: Paired change in grouped MSCE relative to the supervised method in the same family (synthetic minus supervised, $\times 10^3$). Negative is better; † marks a descriptive paired-bootstrap 95% interval entirely below zero. (Bike, Diamonds, Gas Turbine, and Naval.)

Method	Bike	Diamond	Gas Turbine	Naval
SPPI	+1.03	+0.91	+1.92	+1.84
NNM	+0.24	-0.11	+0.03	+0.87
CQR-PPI	-0.79†	-2.69†	-0.30†	-1.93†
RCP-PTFT	-0.78†	-1.53†	-0.99†	-1.02†
RCP-PPI-CV	-2.86†	-2.77†	-0.78†	-2.65†
RCP-Aug	-2.36†	-3.45†	-1.21†	-2.69†
RCP-PPI	-3.24†	-3.79†	-1.41†	-3.88†

Table 5: Paired change in grouped MSCE relative to the supervised method in the same family (synthetic minus supervised, $\times 10^3$). Negative is better; † marks a descriptive paired-bootstrap 95% interval entirely below zero. (SGEMM, Superconductivity, Transcoding, and WEC.)

Method	SGEMM	Supercond.	Transcoding	WEC
SPPI	+0.22	+2.93	+0.70	+2.70
NNM	-0.07	+0.89	-0.05	-0.15
CQR-PPI	-2.85†	-1.61†	-1.53†	+1.01
RCP-PTFT	-0.51†	-0.55	-0.56†	-1.33†
RCP-PPI-CV	-0.90	-2.73†	-0.97†	+1.17
RCP-Aug	-1.42†	-2.65†	-1.25†	-0.79
RCP-PPI	-0.59	-2.95†	-1.55†	+1.29

Table 6: Marginal coverage in the main benchmark (mean ± standard deviation over 30 splits). The target is 0.9. (Bike, Diamonds, Gas Turbine, and Naval.)

Method	Bike	Diamond	Gas Turbine	Naval
Split	0.909±0.023	0.900±0.013	0.898±0.013	0.909±0.023
SPPI	0.892±0.016	0.894±0.008	0.882±0.014	0.887±0.022
NNM	0.904±0.019	0.900±0.011	0.899±0.015	0.900±0.019
CQR	0.904±0.022	0.897±0.011	0.900±0.017	0.906±0.026
CQR-PPI	0.901±0.030	0.898±0.012	0.901±0.015	0.901±0.028
RCP	0.910±0.043	0.905±0.018	0.904±0.024	0.916±0.030
RCP-PTFT	0.910±0.039	0.903±0.019	0.904±0.022	0.912±0.031
RCP-PPI-CV	0.911±0.036	0.906±0.016	0.906±0.023	0.919±0.031
RCP-Aug	0.912±0.036	0.905±0.016	0.905±0.025	0.914±0.031
RCP-PPI	0.910±0.035	0.902±0.014	0.907±0.023	0.915±0.030

Table 7: Marginal coverage in the main benchmark (mean ± standard deviation over 30 splits). The target is 0.9. (SGEMM, Superconductivity, Transcoding, and WEC.)

Method	SGEMM	Supercond.	Transcoding	WEC
Split	0.901±0.013	0.907±0.017	0.897±0.013	0.895±0.016
SPPI	0.896±0.009	0.890±0.022	0.890±0.011	0.882±0.016
NNM	0.901±0.008	0.902±0.022	0.896±0.009	0.895±0.017
CQR	0.903±0.013	0.899±0.020	0.902±0.011	0.895±0.016
CQR-PPI	0.899±0.012	0.902±0.017	0.901±0.013	0.894±0.016
RCP	0.898±0.016	0.905±0.025	0.900±0.017	0.899±0.028
RCP-PTFT	0.900±0.017	0.902±0.029	0.901±0.015	0.901±0.025
RCP-PPI-CV	0.896±0.019	0.904±0.028	0.900±0.015	0.900±0.030
RCP-Aug	0.900±0.017	0.908±0.025	0.901±0.016	0.898±0.028
RCP-PPI	0.894±0.020	0.908±0.026	0.902±0.017	0.901±0.026

Table 8: Mean log volume on the original outcome scale (mean ± standard deviation over 30 splits). Values are comparable only within a dataset; no method produced a full set. (Bike, Diamonds, Gas Turbine, and Naval.)

Method	Bike	Diamond	Gas Turbine	Naval
Split	6.16±0.10	7.64±0.09	4.35±0.14	-7.07±0.17
SPPI	6.08±0.08	7.60±0.06	4.21±0.13	-7.18±0.16
NNM	6.13±0.11	7.65±0.08	4.36±0.17	-7.13±0.17
CQR	5.89±0.10	7.25±0.06	4.18±0.13	-6.88±0.08
CQR-PPI	5.70±0.07	7.09±0.06	4.15±0.12	-6.90±0.09
RCP	6.13±0.18	7.36±0.11	4.39±0.23	-7.04±0.19
RCP-PTFT	6.10±0.15	7.30±0.10	4.34±0.19	-7.07±0.20
RCP-PPI-CV	6.07±0.14	7.29±0.11	4.38±0.22	-7.05±0.22
RCP-Aug	6.08±0.14	7.27±0.09	4.32±0.21	-7.08±0.20
RCP-PPI	6.05±0.12	7.24±0.06	4.37±0.22	-7.09±0.19

Table 9: Mean log volume on the original outcome scale (mean ± standard deviation over 30 splits). Values are comparable only within a dataset; no method produced a full set. (SGEMM, Superconductivity, Transcoding, and WEC.)

Method	SGEMM	Supercond.	Transcoding	WEC
Split	22.46±0.33	4.06±0.09	14.61±0.16	536.00±1.93
SPPI	22.33±0.24	3.98±0.11	14.51±0.13	534.63±1.95
NNM	22.44±0.24	4.04±0.11	14.59±0.10	536.01±2.09
CQR	21.62±0.33	3.68±0.08	13.25±0.17	519.16±2.59
CQR-PPI	20.09±0.24	3.53±0.06	12.71±0.14	516.41±2.25
RCP	21.92±0.35	3.83±0.13	14.22±0.18	530.73±3.70
RCP-PTFT	21.85±0.36	3.78±0.13	14.16±0.16	528.19±3.28
RCP-PPI-CV	21.72±0.32	3.74±0.12	14.09±0.15	525.83±3.51
RCP-Aug	21.78±0.33	3.77±0.12	14.08±0.14	526.05±3.51
RCP-PPI	21.66±0.33	3.74±0.11	14.04±0.13	525.53±3.34

The $K = 10$ raw, debiased, and split-half audits agree with the principal $K = 30$ direction in nearly all dataset–method cells. Complete interval endpoints, win rates, empty-set frequencies, and task-level audit fields are distributed with the results package.

A.5 RCP INTERPRETATION CONTROLS

Direct pooling gives low grouped error on several datasets but assumes that synthetic labels can be trusted without bias correction. RCP-Extra uses true pool outcomes and confirms that additional trusted labels remain a strong resource control. Figure 2 visualizes the matched-permutation contrast, with numerical values in Table 10.

Table 10: Genuine RCP augmentation versus a matched permutation placebo. The placebo preserves the synthetic-score marginal distribution, sample size, architecture, and optimization, but breaks the covariate–score pairing. Entries are paired differences (genuine minus permuted) with 95% bootstrap intervals over 20 common splits. Negative MSCE and positive WSC favor genuine pairing.

Dataset	ΔMSCE$_{30}$ ($\times 10^3$)	ΔMSCE$_{10}$ ($\times 10^3$)	ΔWSC ($\times 100$)
Bike	-1.74 [-2.75, -0.88]	-1.57 [-2.65, -0.72]	1.80 [0.20, 3.40]
Diamond	-4.73 [-5.78, -3.77]	-4.08 [-5.04, -3.14]	10.41 [7.89, 12.78]
Naval	-1.13 [-2.07, -0.34]	-1.06 [-1.86, -0.37]	1.93 [0.11, 3.92]
SGEMM	-1.58 [-2.11, -0.94]	-1.07 [-1.53, -0.55]	4.08 [2.91, 5.33]
Transcoding	-1.14 [-1.56, -0.73]	-0.65 [-0.93, -0.40]	2.91 [1.76, 4.28]

A.6 SENSITIVITY TO LABEL QUALITY AND TRUSTED-LABEL BUDGET

We compare weak, moderate, and oracle labelers on the same ten splits. Budget sweeps vary only nested trusted-data prefixes and keep every other role fixed.

Table 11: Effect of synthetic-label quality. Entries are grouped MSCE $\times 10^3$ for weak/moderate/oracle labelers on the same 10 splits; lower is better. Oracle labels use hidden outcomes and are an upper-bound diagnostic.

Dataset	CQR-PPI	RCP-Aug	**RCP-PPI**
Bike	2.82/2.33/2.02	4.75/4.28/4.21	3.96/3.96/2.99
Diamond	4.48/3.52/0.65	3.57/4.18/3.20	3.85/4.02/1.13
Gas Turbine	2.29/2.04/1.05	2.74/2.74/2.48	3.35/3.00/1.47
Naval	7.65/5.47/2.92	8.38/6.91/6.10	9.20/6.06/4.05
SGEMM	2.59/1.04/0.77	2.51/2.38/2.31	3.95/3.18/2.90
Supercond.	5.64/4.53/2.76	4.52/4.34/3.40	5.57/3.45/1.77
Transcoding	3.19/2.52/1.08	4.16/3.89/3.85	5.58/3.75/1.51
WEC	11.75/10.18/5.97	10.38/7.24/3.04	9.26/9.35/1.69

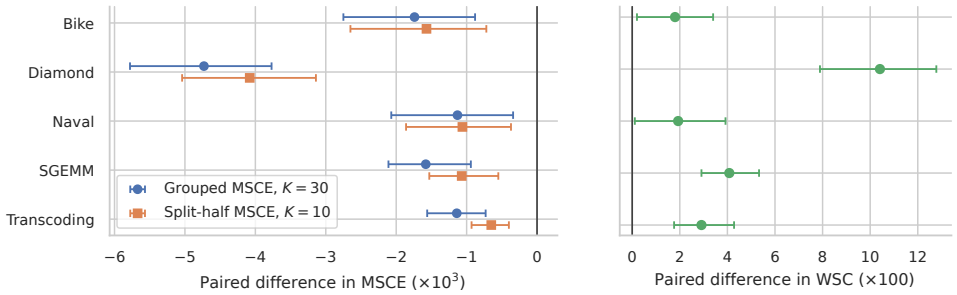

Figure 2: Genuine RCP augmentation versus a matched permutation placebo over 20 common splits. Negative MSCE differences and positive WSC differences favor genuine covariate–score pairing; bars are paired-bootstrap 95% intervals.

Table 12: Trusted-label budget sensitivity. Entries are paired grouped-MSCE changes $\times 10^3$ relative to the supervised method in the same family; negative is better. Ultra-low, low, and moderate use base/conformal proportions of 1%/0.5%, 2%/1%, and 4%/2%, respectively, on the same 10-seed support.

Dataset	Method	Ultra-low	Low	Moderate
Bike	CQR-PPI	-2.79	-0.93	-1.30
Bike	RCP-Aug	-2.94	-3.34	-1.62
Bike	**RCP-PPI**	-3.80	-3.66	-2.49
Diamond	CQR-PPI	-4.24	-2.88	-1.33
Diamond	RCP-Aug	-5.55	-4.14	-2.01
Diamond	**RCP-PPI**	-6.65	-4.30	-1.04
SGEMM	CQR-PPI	-3.77	-2.82	-2.01
SGEMM	RCP-Aug	-2.41	-1.36	-0.56
SGEMM	**RCP-PPI**	-1.49	-0.56	-0.46
Transcoding	CQR-PPI	-1.68	-1.35	-1.29
Transcoding	RCP-Aug	-1.19	-0.88	-0.48
Transcoding	**RCP-PPI**	-2.26	-1.02	-0.47

A.7 Supplementary Experiments

Pinball risk of the score-quantile learner. Proposition 2.1 motivates an intermediate check between the optimization objective and conditional coverage. We evaluate the supervised and unit-PPI score-quantile learners on the same trusted validation folds used for power selection. Table 13 shows lower mean pinball risk under PPI-ERM on every dataset, together with lower across-split variability. This verifies that prediction-powered learning improves the quantile learner before we ask whether the gain survives final conformalization.

Table 13: Intermediate pinball-risk comparison for the RCP score-quantile learner. ERM and unit PPI-ERM are evaluated on the same trusted validation folds used for power selection. We report mean $\pm$ standard deviation over 30 common splits; a positive gain means lower validation risk under PPI-ERM.

Dataset	ERM risk	PPI-ERM risk	Risk reduction	PPI-ERM win rate
Bike	0.1471±0.0585	0.1030±0.0364	0.0441±0.0348	0.93
Diamond	0.0338±0.0169	0.0308±0.0152	0.0030±0.0057	0.63
Gas Turbine	0.1006±0.0405	0.0975±0.0375	0.0030±0.0110	0.57
Naval	0.0930±0.0615	0.0822±0.0301	0.0108±0.0357	0.60
SGEMM	0.0493±0.0166	0.0434±0.0132	0.0059±0.0065	0.73
Supercond.	0.0885±0.0417	0.0754±0.0356	0.0132±0.0164	0.77
Transcoding	0.0794±0.0268	0.0762±0.0265	0.0032±0.0079	0.57
WEC	0.2350±0.0847	0.1649±0.0473	0.0702±0.0623	0.90

Effect of the synthetic-pool size. The main experiments use a large synthetic pool because auxiliary labels are cheaper than trusted outcomes. Table 14 varies only the pool prefix while keeping trusted learning, final calibration, grouping, and evaluation samples fixed. Relative to supervised RCP, even a 5% pool improves all three metrics on the four audited datasets. Further increases help until

the pool-estimation component becomes small, after which performance may plateau or become nonmonotone because the usable covariance and power remain unchanged.

Table 14: Effect of the synthetic-pool size on RCP-PPI. Only the synthetic-pool prefix varies; trusted learning, final calibration, grouping, and evaluation roles are fixed. The 0% row is supervised RCP. Entries are mean ± standard deviation over 10 common splits. MSCE is multiplied by 10^3 (lower is better), WSC is higher-is-better, and L_1-ERT is lower-is-better.

Dataset	Synthetic share	Grouped MSCE	WSC	L_1-ERT
Bike	0%	7.62±3.37	0.818±0.083	0.063±0.016
Bike	5%	4.55±1.98	0.863±0.077	0.045±0.015
Bike	10%	3.66±2.05	0.873±0.067	0.042±0.019
Bike	20%	3.36±1.91	0.873±0.063	0.038±0.021
Bike	40%	3.40±2.18	0.869±0.056	0.038±0.022
Diamond	0%	8.32±3.29	0.713±0.068	0.092±0.010
Diamond	5%	4.32±1.81	0.807±0.062	0.058±0.008
Diamond	10%	4.06±1.11	0.805±0.056	0.057±0.008
Diamond	20%	4.01±1.57	0.803±0.058	0.054±0.008
Diamond	40%	3.85±1.19	0.801±0.049	0.053±0.008
SGEMM	0%	3.74±1.46	0.781±0.042	0.064±0.012
SGEMM	5%	2.11±1.08	0.867±0.044	0.030±0.008
SGEMM	10%	2.73±1.79	0.854±0.040	0.033±0.008
SGEMM	20%	3.41±2.95	0.840±0.050	0.034±0.010
SGEMM	40%	3.57±2.82	0.845±0.053	0.037±0.010
Transcoding	0%	4.77±2.72	0.807±0.048	0.053±0.008
Transcoding	5%	3.89±3.19	0.833±0.072	0.038±0.007
Transcoding	10%	3.91±3.37	0.831±0.071	0.036±0.007
Transcoding	20%	3.91±3.70	0.826±0.059	0.036±0.008
Transcoding	40%	3.89±3.68	0.835±0.059	0.037±0.007

A.8 Realistic LLM-generated supervision on HelpSteer2

After removing exact normalized prompt–response duplicates and one prompt shared between the official train and validation splits, the experiment contains 21,352 rows. The official validation split is held out as test data, and every other role is allocated at the normalized-prompt level, giving zero prompt overlap across base training, LLM pilot, development, final calibration, and test roles. Each response receives five 0–4 human ratings for helpfulness, correctness, coherence, complexity, and verbosity as distinct quality dimensions.

Prompt and response are encoded separately by a frozen `all-MiniLM-L6-v2` model and concatenated into a 768-dimensional feature vector. The five-output base predictor is a $768 \to 128 \to 64 \to 5$ MLP trained on 801 rows. Its nonconformity score is $S(x, y) = \max_{k \le 5} |y_k - \widehat{\mu}_k(x)|$, so one scalar radius defines a five-dimensional box. The score-quantile learner is a $768 \to 64 \to 32 \to 1$ MLP. Within each seed, all methods share initialization. Checkpoint selection uses a separate 100-row human-rated development role, and 400 further human-rated rows are reserved for final conformalization.

The external labelers are the production `deepseek-v4-flash` and `deepseek-v4-pro` routes. They score the same 500 pilot rows at temperature zero in JSON mode, using the five HelpSteer2 rubrics and no in-context examples. All 1,000 calls return valid ratings without retries. For each seed, 100 pilot rows form the paired sample and the remaining 400 form the auxiliary pool. Synthetic-only methods fit only the auxiliary synthetic scores; PPI adds the signed true-minus-synthetic correction on the paired rows. The cost-matched Pro arm keeps the 100 paired calls and uses as many auxiliary Pro calls as fit within the Flash-PPI spend.

Figure 1 reports the main deployed metrics. For the boundary audit, only 13.6% of true pilot scores lie above the target boundary, so an always-below constant-median score obtains 86.4% raw agreement without useful discrimination. We therefore compare each LLM with a shuffled-score reference that preserves its score marginal and above-boundary rate while breaking input-specific alignment. Table 17 reports this comparison. The package includes the executed notebook, rendered report, paper-ready setting, and deterministic figure data.

Table 15: HelpSteer2 downstream results, averaged over 20 paired seeds. Equal-count arms use 100 paired points and 400 auxiliary labels; the cost-matched Pro arm uses 54.8 auxiliary labels on average. Lower MSCE, true pinball risk, and log-volume are better; higher minimum-group coverage is better.

Method	Aux. N	Query cost	True pinball	Coverage	MSCE	Min-group	Log-volume
RCP	0.0	\$0.0000	0.0558	0.8794	0.00665	0.738	-0.426
Flash-SynOnly	400.0	\$0.0422	0.0520	0.8865	0.00648	0.742	-0.370
Flash-PPI	400.0	\$0.0528	0.0611	0.8899	0.00502	0.763	-0.434
Pro-SynOnly	400.0	\$0.1352	0.0507	0.8886	0.00479	0.768	-0.389
Pro-PPI	400.0	\$0.1693	0.0619	0.8937	0.00400	0.773	-0.481
Pro-PPI (cost-matched)	54.8	\$0.0527	0.0721	0.8846	0.00533	0.771	-0.376

Table 16: Paired HelpSteer2 effects relative to RCP. Entries are mean differences with paired-bootstrap 95% intervals. Negative MSCE and log-volume and positive minimum-group coverage favor PPI.

Method	ΔMSCE	Δ minimum-group coverage	Δ log-volume
Flash-PPI	-0.00164 $[-0.00294, -0.00033]$	$+0.0249$ $[-0.0006, +0.0494]$	-0.0078 $[-0.0375, +0.0242]$
Pro-PPI	-0.00265 $[-0.00347, -0.00179]$	$+0.0343$ $[+0.0127, +0.0556]$	-0.0543 $[-0.0909, -0.0168]$
Pro-PPI (cost-matched)	-0.00132 $[-0.00263, -0.00005]$	$+0.0331$ $[+0.0163, +0.0496]$	$+0.0501$ $[+0.0091, +0.0899]$

Table 17: Global rating accuracy and local boundary information on the 500-row HelpSteer2 pilot. Shuffling preserves each synthetic-score marginal while destroying input-specific alignment.

Labeler	Macro MAE	Observed disagree.	Shuffled disagree.	Reduction vs. shuffled	Balanced acc.
Constant median	0.680	0.136	0.136	0.0%	0.500
Flash	0.706	0.134	0.198	32.5%	0.637
Pro	0.846	0.152	0.251	39.5%	0.701

A.9 THEORY DIAGNOSTICS

Table 18 links each principal controlled experiment to a prediction of the post-conformal analysis. These simulations isolate geometry, information mechanisms, and resource scaling; practical evidence comes from the real-data benchmark and permutation experiment.

Table 18: Theory-to-experiment map for the controlled diagnostics.

Prediction	Experiment	Finding
Global shifts and some raw covariance are conformally null	Exact shift check and conformal-null simulation	Sets are unchanged by a constant threshold shift; raw covariance remains positive while projected covariance is zero to numerical precision.
The deployment map is a density-weighted first derivative with a second-order remainder	Population derivative check (Figure 3)	Constant perturbations vanish to machine precision; linear and quadratic remainder slopes are 1.981 and 2.009.
$b = b_{\mathrm{res}} + b_{\mathrm{align}}$ has two sources	Specification $\times$ residual-coupling design	Coupling activates the within-X residual term, misspecification activates between-X alignment, and the coupled misspecified cell contains both.
Exact independent scores give no first-order PPI gain, but certified samples may be pooled	Exact-independent fifth cell	PPI-optimal power is near zero, whereas direct pooling reduces stochastic error when the synthetic law is certified correct.
Power follows a covariance–variance tradeoff	Paired-score oracle and signed-power curves	In the oracle cell $a = b = c$ and the empirical minimum is near $1/(1+\kappa)$; the signed diagnostic recovers the predicted negative optimum.
Learning, pool estimation, and final calibration are separate resources	Joint n, N, m scan, isolated N scan, and calibration-size sweep	Error follows the predicted n^{-1} rate; the isolated $1/N$ coefficient is positive and close to its leading value, while supervision is flat in N; finite-calibration error decreases with $1/m$.

A.10 DIRECT CHECKS OF THE DEPLOYMENT DERIVATIVE AND RESOURCE SCALING

The earlier diagnostics identify null directions and information mechanisms. We add two direct checks of the approximation statements used by the main theory. Both use the controlled score DGP and leave the eight-dataset and HelpSteer2 results unchanged.

Density-weighted deployment derivative. For Corollary 3.5, we evaluate the population conformalization map with 2,048-point Gauss–Legendre quadrature. We perturb the oracle threshold by constant, standardized linear, and standardized quadratic functions at $\epsilon \in \{2^{-1}, \dots, 2^{-8}\}$, recomputing the exact population scalar correction at every ϵ. Constant shifts change the deployed threshold by at most 8.88×10^{-16} and the coverage curve by at most 2.22×10^{-16}. For the two nonconstant directions, the log–log slopes of the $L_2(P_X)$ first-order remainder over the five smallest perturbations are 1.981 and 2.009. Figure 3 therefore checks both the exact null direction and the second-order remainder, rather than only constructing a null covariance in influence space.

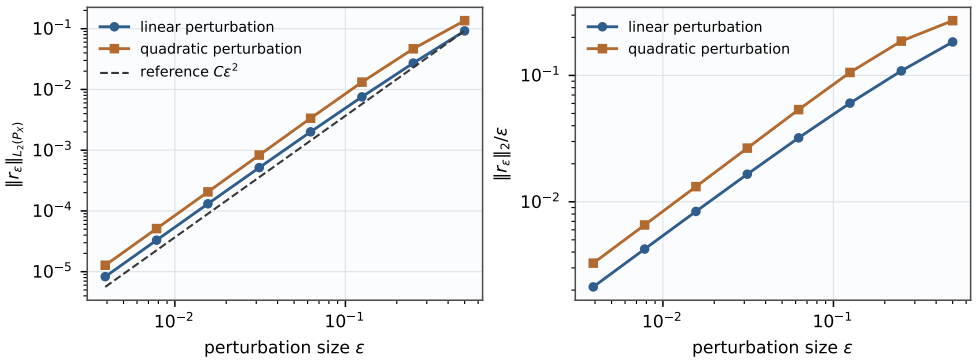

Figure 3: Direct population check of Corollary 3.5. The left panel compares the exact post-conformal first-order remainder with an ϵ^2 reference; the right panel shows that the remainder divided by ϵ vanishes. Constant perturbations are numerical nulls and are omitted from the log-scale curves.

Three-resource scaling. We next use the well-specified residual-coupled cell, fix the population oracle power at 0.2541, and use 30 common seeds per setting. In the joint scan $n \in \{200, 400, 800, 1600\}$ with $N = 4n$ and $m = n$, the finite-calibration deployment-error slopes are -1.159 for oracle power and -1.171 for supervision; the paired-bootstrap interval for the oracle slope is $[-1.364, -0.952]$ and contains the target -1. To isolate the pool term, we fix $n = 400$ and $m = 6400$ and vary $N \in \{400, 800, 1600, 3200, 6400\}$. Regressing population-calibrated error on $1/N$ gives coefficient 0.0240 with interval $[0.00643, 0.04234]$, compared with the leading coefficient 0.0192; the supervised control, which never reads the pool, is exactly flat. Figure 4, together with the earlier calibration-size experiment, supplies separate numerical evidence for the n^{-1}, N^{-1}, and m^{-1} terms. The large-N means are noisy relative to the fixed learning floor, so we interpret signs, rates, and scale rather than pointwise equality.

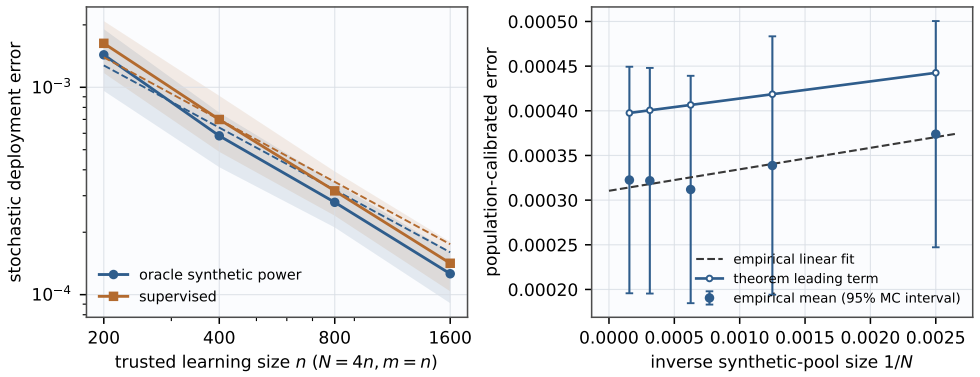

Figure 4: Direct resource-scaling check of Theorem 3.7. Left: finite-calibration deployment error when $N = 4n$ and $m = n$; dashed curves are the leading predictions. Right: population-calibrated oracle-power error against $1/N$ at fixed $n = 400, m = 6400$, with the empirical fit and theorem coefficient. Error bars are 95% Monte Carlo intervals over 30 common seeds.

A.11 Profiled-risk and selection diagnostics

These experiments evaluate consequences of Proposition 3.4 and Proposition B.10; they do not replace fixed unit PPI as the proposed method.

Global-shift audit. For each of 240 dataset–seed–candidate comparisons, we add candidate-specific constants before applying the selection criteria. Raw pinball selection changes in 205 comparisons and the complete raw-risk ordering changes in 235. Cross-calibrated pinball, trusted-only and synthetic-assisted coverage moments, and the final deployed thresholds remain invariant up to 8.88×10^{-16}. This directly checks the quotient geometry in Proposition 3.4; we report it only as an implementation invariant, not as evidence of predictive quality.

Candidate selection. We fix six neural candidates—three training powers $\lambda \in \{0, 0.5, 1\}$ at two checkpoints—and compare raw pinball, cross-calibrated pinball, trusted-only coverage moments, and synthetic-assisted coverage moments. Development data are cross-calibrated, final calibration is untouched, and an Extra-Supervised arm accounts for the additional trusted validation resource. Table 19 and Figure 5 show that the synthetic-assisted selector is worse than fixed unit PPI on Bike, Diamonds, and Transcoding, while its SGEMM interval crosses zero. We therefore retain fixed RCP-PPI and report this study as a negative method result.

Risk–coverage audit on HelpSteer2. Temporary cross-calibration removes global threshold level before measuring pinball risk. Nevertheless, Flash-PPI and Pro-PPI remain worse than supervised RCP by 0.002264 and 0.003030, while their final grouped MSCE improves by 0.001571 and 0.002011. Table 20 therefore exhibits the distinction in Proposition 3.4: risk values and gradient norms need not rank a finite candidate set identically. The reconstructed HelpSteer experiment has prompt-cluster dependence, so these are descriptive paired results rather than a new i.i.d. conformal guarantee.

Quadratic-estimator Monte Carlo. We test Eq. (46) and Eq. (47) in 32 settings with 100,000 repetitions each. The largest absolute bias and variance-formula discrepancies are 2.10 and 2.36 Monte Carlo standard errors. Synthetic correction reduces variance under positive coupling but inflates it under independence or negative coupling, as summarized in Table 21 and Figure 6. Correct estimation therefore does not imply improved model selection.

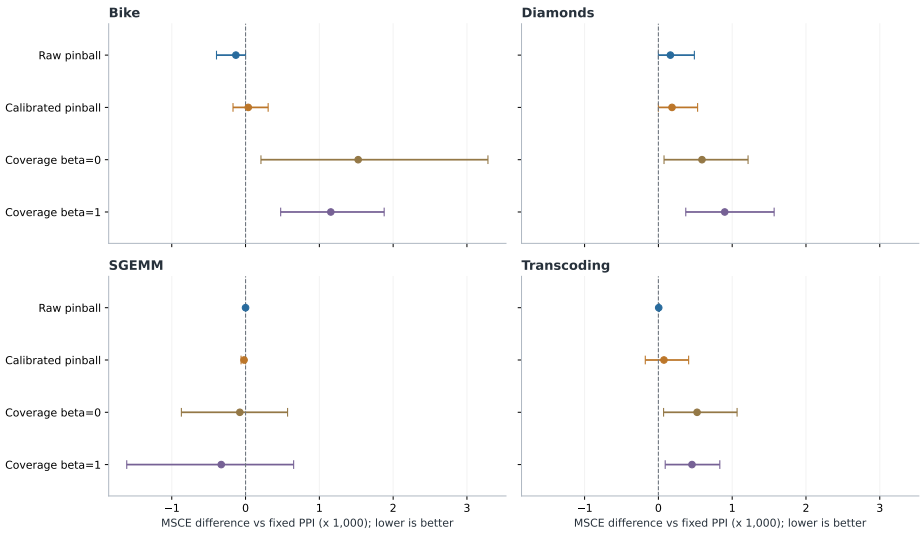

Figure 5: Paired grouped-MSCE changes for the profiled-risk candidate-selection study. Negative values favor the row method. The coverage-based selector does not improve consistently over fixed unit PPI.

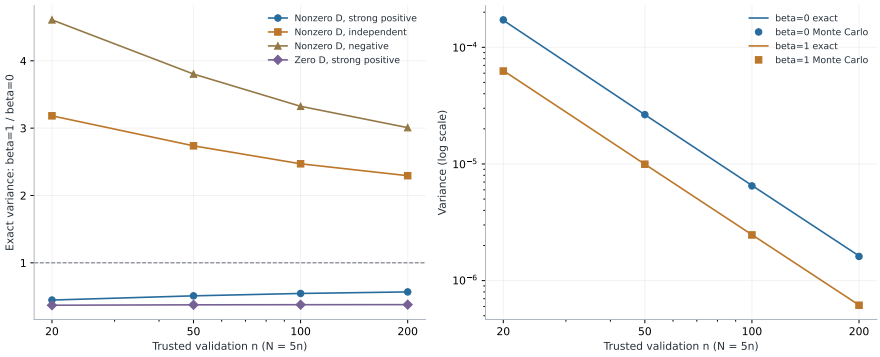

Figure 6: Variance of the unbiased quadratic discrepancy estimator. Synthetic correction helps under positive coupling and hurts when the synthetic moment is independent or negatively coupled.

Table 19: Coverage-moment selection relative to fixed unit PPI. Entries are paired changes in grouped MSCE, multiplied by 10^3; lower is better. Intervals are descriptive paired-bootstrap 95% intervals over 15 common splits.

Dataset	Mean Δ	95% interval	Interpretation
Bike	$+1.155$	$[+0.476, +1.878]$	worse
Diamonds	$+0.899$	$[+0.371, +1.569]$	worse
SGEMM	-0.329	$[-1.610, +0.652]$	inconclusive
Transcoding	$+0.455$	$[+0.092, +0.832]$	worse

Table 20: HelpSteer2 risk–coverage audit. Entries are PPI minus supervised RCP over 20 paired seeds. Lower is better for pinball risk, MSCE, and log volume; higher is better for minimum-group coverage.

Method	Cross-cal. pinball	Group MSCE	Min-group cov.	Log volume
Flash-PPI	$+0.002264$	-0.001571	$+0.01799$	-0.02825
Pro-PPI	$+0.003030$	-0.002011	$+0.02410$	-0.06575

Table 21: Variance ratio of the synthetic-assisted quadratic estimator ($\beta = 1$) to its trusted-only version ($\beta = 0$) across $n \in \{20, 50, 100, 200\}$ and $N = 5n$. Values below one indicate variance reduction.

True–synthetic coupling	Range of variance ratios	Consequence
Positive, nonzero moment	0.45–0.57	reduction
Positive, zero moment	0.37–0.38	reduction
Independent	2.29–3.18	inflation
Negative	3.01–4.61	inflation

B Additional Theoretical Statements

B.1 Regularity and profiled-risk details

The main text states only the deployment-relevant consequences. The profiled correction satisfies

$$D\gamma_q[h] = -\frac{\mathbb{E}\{f_q(X)h(X)\}}{\mathbb{E}f_q(X)}, \tag{31}$$

and the Hessian in Eq. (15) is positive semidefinite because

$$\langle h, \mathcal{H}_q h\rangle = \inf_{c\in\mathbb{R}} \mathbb{E}[f_q(X)\{h(X)+c\}^2] \geq 0. \tag{32}$$

In coverage coordinates $a_0 = f_\star e$, the oracle map is the codimension-one oblique projection

$$(\Pi a_0)(x) = a_0(x) - \frac{f_\star(x)}{\bar{f}_\star}\mathbb{E}\{a_0(X)\}. \tag{33}$$

Assumption B.1 (Local learning and deployment regularity)**.** *Model and curvature.* The threshold family is fixed dimensional and twice differentiable near an interior stationary point θ_0 of the trusted risk, and J is positive definite. *Learning expansion.* Let $P^\dagger$ denote expectation under the common paired/pool synthetic law. Uniformly over λ in a fixed compact set,

$$\widehat{\theta}_\lambda - \theta_0 = -J^{-1}\left[(P_n - P)(\xi - \lambda\xi^\dagger) + \lambda(\widetilde{P}_N - P^\dagger)\widetilde{\xi}\right] + r_\lambda, \tag{34}$$

where $\mathbb{E}\sup_\lambda \|r_\lambda\|^2 = o(n^{-1}+N^{-1})$ and $\mathbb{E}\sup_\lambda \|\widehat{\theta}_\lambda - \theta_0\|^4 = O\{(n^{-1}+N^{-1})^2\}$. *Deployment expansion.* The map from the learned threshold and probability-scale calibration error to $\pi_{\mathcal{D},\lambda}$ has the second-order $L_2(P_X)$ expansion induced by Eq. (20). *Sampling.* The three samples are independent, calibration residuals are continuous, $n/N \to \kappa \in [0,\infty)$, and n/m converges to a positive constant.

B.2 Information mechanisms and consequences

Proposition B.2 (Two sources of useful synthetic information)**.** *For $\mu(X) = \mathbb{E}(\xi \mid X)$ and $\mu^\dagger(X) = \mathbb{E}(\xi^\dagger \mid X)$,*

$$b = \underbrace{\operatorname{tr}\big[G_c\mathbb{E}\{\operatorname{Cov}(\xi,\xi^\dagger \mid X)\}\big]}_{b_{\mathrm{res}}} + \underbrace{\operatorname{tr}\big[G_c\operatorname{Cov}\{\mu(X),\mu^\dagger(X)\}\big]}_{b_{\mathrm{align}}}. \tag{35}$$

Exact specification removes b_{align}, whereas $S^\dagger \perp S \mid X$ removes b_{res}.

Corollary B.3 (No first-order PPI gain under exact independent synthetic scores)**.** *Under exact specification, if $S^\dagger \perp S \mid X$, then $b = 0$ and $\lambda = 0$ is a first-order optimum over nonnegative powers, uniquely so when $c > 0$.*

This result is specific to the matched bias-robust correction in Eq. (6); certified-correct synthetic samples may still be pooled under Proposition B.7.

Corollary B.4 (Exact score-quantile specification)**.** *If $q_0 = q^\star$, then $\pi_0 \equiv \tau$ and*

$$\mathbb{E}\,\mathrm{MSCE}_{\mathcal{D}} = L(\lambda) + C_{\mathrm{cal}}\sigma^2_{m,\tau} + o(n^{-1}+N^{-1}+m^{-1}), \tag{36}$$

$$\mathbb{E}\,\mathrm{HET}_{\mathcal{D}} = L(\lambda) + (C_{\mathrm{cal}}-1)\sigma^2_{m,\tau} + o(n^{-1}+N^{-1}+m^{-1}), \tag{37}$$

where $C_{\mathrm{cal}} - 1 = \operatorname{Var}\{f_\star(X)\}/\bar{f}^2_\star$.

Thus boundary-density heterogeneity is the first-order multiplier that turns scalar calibration noise into conditional heterogeneity. Under misspecification, Theorem 3.7 controls stochastic error around π_0; total MSCE also includes approximation and interaction.

Lemma B.5 (Reparameterization invariance)**.** *Under any locally invertible differentiable reparameterization, the random function $\mathsf{B}\xi$ is unchanged. Consequently, a, b, c, the oracle scalar power, and its leading regret are invariant, although their coordinate matrices change.*

Corollary B.6 (Conformal null information)**.** *For centered influences,*

$$\operatorname{tr}\{G_c\operatorname{Cov}(\xi,\xi^\dagger)\} = \mathbb{E}\langle \mathsf{B}\xi, \mathsf{B}\xi^\dagger\rangle_{L_2(P_X)}. \tag{38}$$

Raw covariance can therefore be nonzero while deployed covariance is zero, particularly when the shared information lies only in a global threshold-shift direction.

Proposition B.7 (Certified pooling versus bias-robust PPI)**.** *Under exact specification and a certified synthetic joint law equal to the target law, direct pooling has leading post-conformal learning error*

$$\frac{a}{n+N} + o(n^{-1} + N^{-1}). \tag{39}$$

For a local synthetic-law path satisfying, uniformly for bounded h,

$$\sqrt{n}\,\nabla_\theta \mathcal{R}_{Q_n}(q_{\theta_0 + h/\sqrt{n}}) = \delta + Jh + o(1), \tag{40}$$

with matching derivative and variance limits, pooling incurs post-conformal squared bias

$$\frac{\delta^\top G_c \delta}{n(1+\kappa)^2} + o(n^{-1}) \tag{41}$$

and leading risk

$$\frac{a}{n+N} + \frac{\Delta^2}{n(1+\kappa)^2}, \qquad \Delta^2 = \delta^\top G_c \delta. \tag{42}$$

It improves on a/n at this local order precisely when $\Delta^2 < a(1+\kappa)$. Matched paired–pool PPI remains centered at the trusted target because its two synthetic expectations cancel.

Corollary B.8 (Conservative estimable power)**.** *Let $\widehat{b} - b = o_p(t_n)$, $\widehat{c} - c = o_p(1)$, and $\widehat{\kappa} \to_p \kappa$, with $t_n \to 0$ and $\sqrt{n}\,t_n \to \infty$. For $\Lambda = [0, \lambda_{\max}]$,*

$$\widehat{\lambda}_{\mathrm{ad}} = \Pi_\Lambda \left\{ \frac{(\widehat{b} - t_n)_+}{(1+\widehat{\kappa})\max(\widehat{c}, \epsilon_n)} \right\}, \qquad \epsilon_n \downarrow 0. \tag{43}$$

If $b \le 0$ and $c > 0$, the rule is asymptotically inactive. If $b, c > 0$ and the oracle is interior, it converges to $\lambda^\star$ and has $o_p(n^{-1})$ leading regret.

Corollary B.9 (Oracle trusted-label allocation)**.** *For proportional $N = n/\kappa$, trusted budget $L = n + m$, learning constant $C_{\mathrm{learn}} = \min_\lambda\{a - 2b\lambda + (1+\kappa)c\lambda^2\}$, and $C_{\mathrm{final}} = \tau(1-\tau)C_{\mathrm{cal}}$,*

$$n^\star = L\frac{\sqrt{C_{\mathrm{learn}}}}{\sqrt{C_{\mathrm{learn}}} + \sqrt{C_{\mathrm{final}}}}, \qquad m^\star = L\frac{\sqrt{C_{\mathrm{final}}}}{\sqrt{C_{\mathrm{learn}}} + \sqrt{C_{\mathrm{final}}}}. \tag{44}$$

B.3 Finite-resolution post-conformal evaluation

The profiled-risk identity concerns the full conditional-coverage function. For evaluation with finitely many groups or bounded features, fix $\varphi(X) \in \mathbb{R}^d$ and a candidate threshold $q + \gamma$ independently of the evaluation samples. Let

$$Z = \mathbb{1}\{S \le q(X) + \gamma\} - \tau, \qquad \mu_\varphi = \mathbb{E}[Z\varphi(X)], \qquad D_\varphi = \|\mu_\varphi\|^2. \tag{45}$$

For normalized group indicators, D_φ is population grouped MSCE; for general features it measures only the projected component of conditional error.

Let $Z_i^\dagger$ and $\widetilde{Z}_j$ be paired and pool synthetic boundary indicators at the same threshold. For a fixed evaluation weight β, define

$$A_i = \varphi(X_i)(Z_i - \beta Z_i^\dagger), \qquad B_j = \varphi(\widetilde{X}_j)\widetilde{Z}_j,$$

and let $\widehat{\Sigma}_A, \widehat{\Sigma}_B$ be the unbiased sample covariance matrices.

Proposition B.10 (Unbiased synthetic-assisted quadratic discrepancy)**.** *If the paired and pool synthetic laws match and the two evaluation samples are independent, then*

$$\widehat{D}_\beta = \|\bar{A} + \beta\bar{B}\|^2 - \frac{\mathrm{tr}(\widehat{\Sigma}_A)}{n_v} - \frac{\beta^2\,\mathrm{tr}(\widehat{\Sigma}_B)}{N_v} \tag{46}$$

is unbiased for D_φ. Writing $\Sigma_A = \mathrm{Var}(A_i)$ and $\Sigma_B = \mathrm{Var}(B_j)$, its exact variance is

$$\begin{aligned}\mathrm{Var}(\widehat{D}_\beta) = 4\mu_\varphi^\top\left(\frac{\Sigma_A}{n_v} + \frac{\beta^2\Sigma_B}{N_v}\right)\mu_\varphi + \frac{2\,\mathrm{tr}(\Sigma_A^2)}{n_v(n_v-1)} + \frac{2\beta^4\,\mathrm{tr}(\Sigma_B^2)}{N_v(N_v-1)} \\ + \frac{4\beta^2\,\mathrm{tr}(\Sigma_A\Sigma_B)}{n_v N_v}.\end{aligned} \tag{47}$$

When $\mu_\varphi = 0$, the first-order term vanishes and evaluation is intrinsically second order. Synthetic correction can therefore reduce or increase selection noise; the training-stage power rule in Corollary 3.8 does not transfer unchanged to β.

C TECHNICAL RESULTS AND PROOFS

C.1 PINBALL RISK AND PRE-CONFORMAL COVERAGE

Proof of Proposition 2.1. Fix x and write $r_x(t) = \mathbb{E}[\rho_\tau(S-t) \mid X = x]$. The integral identity for the pinball loss gives

$$r_x(t) - r_x(q^\star(x)) = \int_{q^\star(x)}^{t} \{F_x(u) - \tau\}\, du,$$

where an integral with reversed limits has the usual sign. Suppose first that $t \geq q^\star(x)$ and set $\delta = F_x(t) - \tau \geq 0$. Since F_x is nondecreasing and L_F-Lipschitz,

$$F_x(u) - \tau \geq \{\delta - L_F(t-u)\}_+$$

for $u \in [q^\star(x), t]$. Integrating over the support of the right-hand side gives

$$r_x(t) - r_x(q^\star(x)) \geq \frac{\delta^2}{2L_F}.$$

The case $t < q^\star(x)$ is identical after setting $\delta = \tau - F_x(t)$ and reversing the integral. Hence

$$\{F_x(t) - \tau\}^2 \leq 2L_F\{r_x(t) - r_x(q^\star(x))\}.$$

Substituting $t = q(x)$ and integrating over P_X proves the claim. □

C.2 EXACT MSCE IDENTITIES AND PROBABILITY-SCALE CALIBRATION

Proof of Proposition 3.2. For a fixed deployment, add and subtract $U_{\mathcal{D}} = \mathbb{E}_X \pi_{\mathcal{D}}(X)$ and use the orthogonality of a centered random variable and a constant:

$$\mathbb{E}_X[(\pi_{\mathcal{D}}(X) - \tau)^2] = \mathrm{Var}_X\{\pi_{\mathcal{D}}(X)\} + (U_{\mathcal{D}} - \tau)^2.$$

Conditional on $\widehat{q}$, let $R = S - \widehat{q}(X)$ have continuous CDF F_R. Since $U_{\mathcal{D}} = F_R(\widehat{\gamma})$ and the calibration residuals are i.i.d. from F_R, the probability integral transform gives $F_R(\widehat{\gamma}) \sim \mathrm{Beta}(k, m+1-k)$. Its mean and variance yield Eq. (9). □

Proof of Eq. (10). For each fixed x, apply the bias–variance identity to $\pi_{\mathcal{D}}(x)$ around τ; integrating the resulting equality over P_X proves the claim. □

Proof of Proposition 3.3. Replacing q by $q + c$ shifts every calibration residual by $-c$, hence shifts its kth order statistic by $-c$. The sum $q(x) + \widehat{\gamma}(q)$ and therefore the prediction set are unchanged. □

Proof of Proposition 3.1. Conditional on $(\mathcal{D}_L, \mathcal{D}_U^X)$ and the fixed fitted objects and the auxiliary learning randomness, $\widehat{\lambda}$ and $\widehat{q}$ are fixed, while the m calibration residuals and the test residual are exchangeable. Under continuity the rank is uniform. With ties, the conservative kth order statistic still gives the displayed lower bound; randomized ranks can recover exact rank validity but are not used for the Beta identity in Proposition 3.2. The infinite-set convention covers $k = m+1$. □

Proposition C.1 (Exact transport and a finite-sample bridge)**.** *Fix a measurable q independent of calibration and let $R = S - q(X)$ have a continuous strictly increasing CDF F_R. Define*

$$U_{(k)} = F_R(\widehat{\gamma}), \quad g_x(u) = F_{R|X=x}\{F_R^{-1}(u)\}, \quad u_0 = F_R(0).$$

Then $U_{(k)} \sim \mathrm{Beta}(k, m+1-k)$ and

$$\pi_{\mathcal{D}}(x) = g_x(U_{(k)}), \qquad F_{S|X}(q(x) \mid x) = g_x(u_0).$$

If $\sup_r f_{R|X=x}(r)/f_R(r) \leq \Lambda$, then, with $v_m = \mathrm{Var}(U_{(k)})$ and $\tau_m = \mathbb{E}U_{(k)}$,

$$\left|\mathbb{E}_{\mathcal{D}_C}\pi_{\mathcal{D}}(x) - F_{S|X}(q(x) \mid x)\right| \leq \Lambda\{|\tau_m - u_0| + \sqrt{v_m}\}, \tag{48}$$

$$\mathbb{E}_{\mathcal{D}_C}[(\pi_{\mathcal{D}}(x) - \tau)^2] \leq 2\{F_{S|X}(q(x) \mid x) - \tau\}^2 + 2\Lambda^2\{(\tau_m - u_0)^2 + v_m\}. \tag{49}$$

Proof. The probability integral transform gives the Beta law and the displayed identities. The density-ratio bound implies $g_x'(u) = f_{R|X=x}(F_R^{-1}(u))/f_R(F_R^{-1}(u)) \leq \Lambda$, hence g_x is Λ-Lipschitz. Jensen's inequality and the Beta moments prove Eq. (48); applying $(a+b)^2 \leq 2a^2 + 2b^2$ proves Eq. (49). □

C.3 Profiled risk and post-conformal geometry

Proof of Proposition 3.4. Shift invariance follows by changing variables in the scalar infimum: $\inf_\gamma \mathcal{R}(q+c+\gamma) = \inf_{\gamma'} \mathcal{R}(q+\gamma')$. For fixed (q,γ), the directional derivative of the pinball risk is

$$D\mathcal{R}(q+\gamma)[h] = \mathbb{E}[\{F_X(q(X)+\gamma) - \tau\}h(X)].$$

At the locally unique minimizer γ_q, the scalar first-order condition is $\mathbb{E}\{\pi_q(X) - \tau\} = 0$. The envelope theorem therefore removes the derivative of γ_q and gives Eq. (13). The $L_2(P_X)$ gradient and Eq. (14) follow immediately.

For the second derivative, differentiate the calibration equation $\mathbb{E}F_X\{q(X)+\gamma_q\} = \tau$ along $q + th$ at $t = 0$. This gives

$$0 = \mathbb{E}[f_q(X)\{h(X) + D\gamma_q[h]\}],$$

which proves Eq. (31). Differentiating the gradient $\pi_q - \tau$ then yields Eq. (15). Finally,

$$\langle h, \mathcal{H}_q h\rangle = \mathbb{E}(f_q h^2) - \frac{\{\mathbb{E}(f_q h)\}^2}{\mathbb{E}f_q} = \inf_{c\in\mathbb{R}} \mathbb{E}[f_q(X)\{h(X)+c\}^2],$$

where the minimizing constant is $c = -\mathbb{E}(f_q h)/\mathbb{E}f_q$. This proves positive semidefiniteness. □

Proof of Corollary 3.5. Write $\gamma_{e,v} = \Gamma(q^\star + e, \tau + v)$ and $z_{e,v}(x) = e(x) + \gamma_{e,v}$. A Taylor expansion of the marginal constraint, uniform in the stated neighborhood, gives

$$v = \mathbb{E}[f_\star(X) z_{e,v}(X)] + O\{(\|e\|_\infty + |\gamma_{e,v}|)^2\}.$$

Local uniqueness and $\bar{f}_\star > 0$ imply $|\gamma_{e,v}| \lesssim \|e\|_\infty + |v|$. Solving the preceding equation yields

$$\gamma_{e,v} = -\frac{\mathbb{E}(f_\star e)}{\bar{f}_\star} + \frac{v}{\bar{f}_\star} + O\{(\|e\|_\infty + |v|)^2\}.$$

A pointwise Taylor expansion of F_x at $q^\star(x)$, together with the uniform bound on the density derivative, gives an $L_2(P_X)$ remainder of order $(\|e\|_\infty + |v|)^2$. Substitution proves Eqs. (17)–(18). Equation (33) follows by setting $a_0 = f_\star e$. □

For a parametric path $e_v(x) = d(x)^\top v$, Corollary 3.5 gives $D\mathcal{P}_{(q_0,\tau)}[e_v, 0] = f_c d_c^\top v$. This proves the finite-dimensional derivative asserted around Eq. (19) and shows that G_c is positive semidefinite. It can be singular because constant-shift directions are removed; all correlation statements therefore use the induced semi-inner product, or equivalently the quotient by $\ker(G_c)$.

C.4 Stationary-target PPI asymptotics

Assumption C.2 (Stationary local model and uniform asymptotic linearity)**.** Fix a compact admissible power set Λ. Conditional on the fixed fitted objects, the three splits and test pair are mutually independent, $(X, S^\dagger) \overset{d}{=} (\widetilde{X}, \widetilde{S})$, $n/N \to \kappa \in [0,\infty)$, and $n/m \to \nu \in (0,\infty)$. Calibration residuals have a continuous marginal CDF. Ties may be handled conservatively for Proposition 3.1, but the Beta expansion below is asserted only under continuity.

The parameter space contains a neighborhood of an interior stationary point θ_0; J is positive definite; q_θ is twice continuously differentiable with integrable fourth-moment envelopes; and the true and synthetic boundary densities are locally regular. Uniformly over $\lambda \in \Lambda$,

$$\widehat{\theta}_\lambda - \theta_0 = -J^{-1} Z_\lambda + R_\lambda, \qquad Z_\lambda = (P_n - P)(\xi - \lambda\xi^\dagger) + \lambda(\widetilde{P}_N - P^\dagger)\widetilde{\xi}, \tag{50}$$

with

$$\frac{\mathbb{E}[\sup_{\lambda\in\Lambda}\|R_\lambda\|^2]}{n^{-1}+N^{-1}} \to 0, \qquad \mathbb{E}\left[\sup_{\lambda\in\Lambda}\|\widehat{\theta}_\lambda - \theta_0\|^4\right] = O\{(n^{-1}+N^{-1})^2\}. \tag{51}$$

Remark C.3 (Proof level of the learning expansion). Assumption C.2 is a high-level uniform asymptotic-linearity condition. The contribution of Theorem 3.7 is the downstream post-conformal geometry and resource decomposition conditional on such a learning expansion, not a claim that arbitrary signed ERM automatically satisfies it.

For the pinball risk, write

$$g_\theta(x,s) = \{\mathbb{1}(s \le q_\theta(x)) - \tau\}\nabla_\theta q_\theta(x).$$

Proposition C.4 (A sufficient fixed-dimensional one-step construction)**.** *Suppose $q_\theta(x) = d(x)^\top\theta$ has fixed dimension and bounded fourth-moment feature envelope; true and synthetic conditional CDFs are continuously differentiable in a uniform local tube with Lipschitz densities; J is positive definite; $\widetilde{\theta}$ is a root-n trusted-only pilot; and $\widehat{J} \to_p J$. Define*

$$\widehat{\theta}^{\mathrm{os}}_\lambda = \widetilde{\theta} - \widehat{J}^{-1}\left[P_n g_{\widetilde{\theta}}(X,S) + \lambda\{\widetilde{P}_N g_{\widetilde{\theta}}(\widetilde{X},\widetilde{S}) - P_n g_{\widetilde{\theta}}(X,S^\dagger)\}\right]. \tag{52}$$

Under the standard fixed-dimensional quantile-score stochastic-equicontinuity conditions, Eq. (50) *holds uniformly on compact Λ with $R_\lambda = o_p(n^{-1/2} + N^{-1/2})$. Bounded higher moments give the expected remainder controls in Eq.* (51)*.*

Proof. Expand the one-step score at θ_0. The trusted population score vanishes by stationarity, while equality in law of paired and pool synthetic observations cancels their population means. Uniform stochastic equicontinuity controls replacement of $\widetilde{\theta}$ by θ_0; consistency of $\widehat{J}$ and the root-n pilot make the remaining terms second order. Compactness of Λ makes the affine dependence on λ uniform. Standard fourth-moment maximal inequalities yield the stated expected controls. □

Proposition C.5 (Stability boundary of signed pinball ERM)**.** *For fixed scores $s, s^\dagger$, let $g_\lambda(t) = \rho_\tau(s-t) - \lambda\rho_\tau(s^\dagger - t)$. If $\lambda > 1$, then $g_\lambda(t) \to -\infty$ as $t \to \pm\infty$; if $\lambda = 1$, both tails converge to finite constants and g_1 is not coercive. Thus the local theorem concerns a stable root or one-step update; unrestricted rich-class signed ERM may be ill posed.*

Proof. For t above both scores, $g_\lambda(t) = (1-\tau)\{(1-\lambda)t - s + \lambda s^\dagger\}$; below both scores, $g_\lambda(t) = \tau\{s - \lambda s^\dagger - (1-\lambda)t\}$. The conclusions follow directly. □

Lemma C.6 (Second-moment consequence)**.** *Under Assumption C.2, uniformly over $\lambda \in \Lambda$,*

$$\begin{aligned}\mathbb{E}[(\widehat{\theta}_\lambda - \theta_0)(\widehat{\theta}_\lambda - \theta_0)^\top] = \frac{1}{n}J^{-1}\operatorname{Var}(\xi - \lambda\xi^\dagger)J^{-1} \\ + \frac{\lambda^2}{N}J^{-1}\operatorname{Var}(\widetilde{\xi})J^{-1} + o(n^{-1} + N^{-1}).\end{aligned} \tag{53}$$

Proof. The two empirical terms are independent and centered. Expand Eq. (50); the L_2 remainder bound and Cauchy–Schwarz make every remainder and cross term negligible uniformly over Λ. □

Proof of Theorem 3.7. Conditional on $\widehat{\theta}_\lambda$, the probability-scale calibration variable $U = F_{\widehat{\theta}_\lambda}(\widehat{\gamma})$ is $\mathrm{Beta}(k, m+1-k)$ and independent of the learning samples. Assumption B.1 and Corollary 3.5 give

$$\pi_{\mathcal{D},\lambda}(X) - \pi_0(X) = f_c(X)d_c(X)^\top(\widehat{\theta}_\lambda - \theta_0) + \frac{f_c(X)}{\mathbb{E}f_c(X)}(U-\tau) + r_X.$$

The fourth-moment controls localize the expansion with complement probability $o(n^{-1} + N^{-1} + m^{-1})$; bounded coverage and uniform integrability handle the complement. Taking the expected squared $L_2(P_X)$ norm, applying Lemma C.6, and using independence yields Eq. (24). Centering over the procedure replaces $\mathbb{E}(U-\tau)^2$ by $\operatorname{Var}(U)$ and gives Eq. (25). Corollary B.4 follows from $\pi_0 = \tau$ and the exact MSCE decomposition in Proposition 3.2. □

Proof of Lemma B.5 and Corollary B.6. For a local linear change of coordinates $\theta = A\eta$, one has $d_\eta = A^\top d_\theta$, $J_\eta = A^\top J_\theta A$, and $\xi_\eta = A^\top\xi_\theta$. Substitution in Eq. (20) gives $\mathsf{B}_\eta\xi_\eta = \mathsf{B}_\theta\xi_\theta$. A differentiable reparameterization reduces to its invertible Jacobian locally. The Hessian transforms by congruence because the risk gradient vanishes at the stationary target, so the second-coordinate-derivative term from a nonlinear change of variables drops out. Equation (38) follows by polarization and $G_c = (\mathsf{B})^*\mathsf{B}$. A global-shift score Jw is in the kernel because $d_c(X)^\top w = 0$. □

C.5 Information boundary, power, and allocation

Proof of Proposition B.2 and Corollary B.3. The law of total covariance decomposes $\mathrm{Cov}(\xi, \xi^\dagger)$ into $\mathbb{E}\{\mathrm{Cov}(\xi, \xi^\dagger \mid X)\}$ and $\mathrm{Cov}\{\mu(X), \mu^\dagger(X)\}$. Applying $M \mapsto \mathrm{tr}(G_c M)$ proves Eq. (35). Under exact specification, $\mu(X) = \{F_X(q^\star(X)) - \tau\} d(X) = 0$. Conditional independence also makes the conditional covariance zero, so $b = 0$ and Eq. (23) is minimized at zero over nonnegative powers. □

Proof of Proposition B.7. Under certified correctness, trusted and synthetic estimating scores are independent and identically distributed. The pooled score average has covariance $\mathrm{Var}(\xi)/(n+N)$, and the post-conformal map gives Eq. (39). In the robust model, conditional independence invokes the zero-power conclusion of Corollary B.3.

For the local path, Eq. (40) supplies both the first-order score displacement and the common limiting derivative J; the variance assumption supplies the pooled covariance limit. The synthetic fraction is $N/(n+N) = 1/(1+\kappa) + o(1)$, so the pooled stationary point is shifted by $-J^{-1}\delta/\{(1+\kappa)\sqrt{n}\} + o(n^{-1/2})$. Applying the post-conformal map gives Eq. (41). Adding the pooled variance $a/(n+N)$ and comparing with a/n gives the threshold. If paired and pool synthetic observations follow the same Q_n, their population score contributions cancel in Eq. (6). □

Proof of Corollary 3.8. Expanding the learning variance in Theorem 3.7 gives Eq. (23). Differentiation and clipping yield Eq. (27). Cauchy–Schwarz in the semi-inner product induced by G_c gives $|b| \le \sqrt{ac}$; substitution gives the minimum, and completing the square gives Eq. (28). □

Corollary C.7 (Paired-score oracle check)**.** *If $\xi^\dagger = \xi$ almost surely and the pool has the same boundary-score law, then $a = b = c$ and*

$$\lambda^\star = \frac{1}{1+\kappa}, \qquad L(\lambda^\star) = \frac{a}{n+N}. \tag{54}$$

C.6 Cross-fitted estimation of projected boundary information

Partition the paired sample into a fixed number of folds. On each complement, fit a trusted-only pilot, nearby quantile heads for conditional boundary density, the derivative map, and a positive-semidefinite estimate of G_c. Evaluate both paired boundary scores and the entire synthetic pool using that fold-specific pilot. Center within fold and average the resulting bilinear forms to obtain $\widehat{b}$; use fold-specific pool scores for $\widehat{c}$.

Proposition C.8 (Consistency of cross-fitted projected constants)**.** *If the fold-specific threshold, derivative, density, and geometry estimates converge in the corresponding L_2/L_4 and operator norms, the number of folds is fixed, and relevant fourth moments are uniformly bounded, then $\widehat{a} \to_p a$, $\widehat{b} \to_p b$, and $\widehat{c} \to_p c$.*

Proof. Conditional on each off-fold pilot, validation observations and the independent pool obey their target laws. Conditional laws of large numbers, fold-mean consistency, indicator convergence away from vanishing boundary neighborhoods, and uniform integrability yield the required population bilinear forms and consistency. □

Proof of Corollary B.8. If $b < 0$, consistency places $\widehat{b} - t_n$ below zero with probability tending to one. If $b = 0$, the stronger $o_p(t_n)$ condition gives the same conclusion. If $b, c > 0$, consistency and $t_n, \epsilon_n \to 0$ give convergence of the unclipped ratio to $b/\{(1+\kappa)c\}$; interiority removes clipping. Equation (28) then gives $o_p(n^{-1})$ regret. □

Proposition C.9 (Boundary disagreement is a coarse upper bound)**.** *At exact specification, let $D_\tau = \mathbb{1}\{\mathbb{1}(S \le q^\star(X)) \ne \mathbb{1}(S^\dagger \le q^\star(X))\}$. Then*

$$\mathrm{tr}\{G_c \,\mathrm{Var}(\xi - \xi^\dagger)\} \le \mathbb{E}[D_\tau d(X)^\top G_c d(X)]. \tag{55}$$

This diagnostic does not identify the sign of b.

Proof. The score difference is the boundary-indicator difference times $d(X)$. For positive-semidefinite G_c, $\mathrm{tr}\{G_c \,\mathrm{Var}(V)\} \le \mathbb{E}(V^\top G_c V)$. □

Proof of Corollary B.9. Minimize $C_{\text{learn}}/n + C_{\text{final}}/(L-n)$. The first-order condition and strict convexity yield Eq. (44). □

Proposition C.10 (X-only information ceiling)**.** *In a pseudo-true regime with $S^\dagger \perp S \mid X$, let $\mu(X) = \mathbb{E}(\xi \mid X)$ and $a_X = \operatorname{tr}\{G_c \operatorname{Var}(\mu(X))\}$. Then*

$$a = a_X + \operatorname{tr}[G_c \mathbb{E}\{\operatorname{Var}(\xi \mid X)\}], \qquad \frac{b^2}{c} \le a_X, \tag{56}$$

with the ratio zero when $c = 0$. Thus a scalar X-only control can remove at most the between-X part of the leading learning constant; a finite pool further scales the attainable reduction by $1/(1+\kappa)$.

Proof. Conditional independence gives $\operatorname{Cov}(\xi, \xi^\dagger) = \operatorname{Cov}\{\mu(X), \mathbb{E}(\xi^\dagger \mid X)\}$. Apply the law of total variance and Cauchy–Schwarz after mapping both variables by $(G_c)^{1/2}$. □

C.7 Synthetic-law mismatch

Proposition C.11 (Bias from paired–pool mismatch)**.** *Let $A_c = \mathbb{E}[f_c(X)^2 d_c(X) d_c(X)^\top]$, so $G_c = J^{-1} A_c J^{-1}$. Define*

$$M_t(\theta) = \mathbb{E}_{\text{pool},t} g_\theta(\widetilde{X}, \widetilde{S}) - \mathbb{E}_{\text{paired},t} g_\theta(X, S^\dagger), \qquad \mu_{\dagger,t} = M_t(\theta_0) \to 0.$$

Assume M_t is differentiable at θ_0, its derivative is locally equicontinuous, and $K_{\lambda,t} = J + \lambda \dot{M}_t(\theta_0)$ remains nonsingular. In the well-specified local model,

$$\theta_{\lambda,t} - \theta_0 = -\lambda K_{\lambda,t}^{-1} \mu_{\dagger,t} + o(\|\mu_{\dagger,t}\|), \tag{57}$$

and the post-conformal squared bias is

$$\lambda^2 \mu_{\dagger,t}^\top K_{\lambda,t}^{-\top} A_c K_{\lambda,t}^{-1} \mu_{\dagger,t} + o(\|\mu_{\dagger,t}\|^2). \tag{58}$$

If additionally $\|\dot{M}_t(\theta_0)\| \to 0$, this reduces to $\lambda^2 \mu_{\dagger,t}^\top G_c \mu_{\dagger,t} + o(\|\mu_{\dagger,t}\|^2)$.

Proof. The expected PPI gradient is $\nabla \mathcal{R}_P(\theta) + \lambda M_t(\theta)$. A local implicit expansion around θ_0 has derivative $K_{\lambda,t}$ and value $\lambda \mu_{\dagger,t}$, yielding Eq. (57). The parameter perturbation has deployed squared norm $v^\top A_c v$; substitution gives Eq. (58). The final reduction follows from $K_{\lambda,t} \to J$. □

For a fixed nonzero paired–pool mismatch, the population objective generally targets a shifted parameter and increasing n or N does not restore the trusted target. Proposition C.11 does not claim that its local remainder formula describes arbitrarily large mismatch; importance weighting or an explicit transport model is then required.

C.8 Quadratic post-conformal discrepancy

Proof of Proposition B.10. Matched paired and pool synthetic laws imply

$$\mathbb{E}(\bar{A} + \beta \bar{B}) = \mathbb{E}[\varphi(X)\{Z - \beta Z^\dagger\}] + \beta \mathbb{E}[\varphi(X) Z^\dagger] = \mu_\varphi.$$

For any i.i.d. vector sample $V_1, \ldots, V_r$ with mean μ_V, the identity

$$\|\bar{V}\|^2 - \frac{\operatorname{tr}(\widehat{\Sigma}_V)}{r} = \frac{1}{r(r-1)} \sum_{i \neq j} V_i^\top V_j$$

shows that the corrected squared mean is unbiased for $\|\mu_V\|^2$. Applying this identity to the A and B samples and retaining the independent cross term $2\beta \bar{A}^\top \bar{B}$ proves Eq. (46).

For the variance, use the Hoeffding decomposition of each second-order inner-product U-statistic. Its linear projection has variance $4\mu_\varphi^\top \Sigma_A \mu_\varphi / n_v$ for the A sample and $4\beta^2 \mu_\varphi^\top \Sigma_B \mu_\varphi / N_v$ for the B sample. The two degenerate components contribute $2 \operatorname{tr}(\Sigma_A^2)/\{n_v(n_v-1)\}$ and $2\beta^4 \operatorname{tr}(\Sigma_B^2)/\{N_v(N_v-1)\}$. Independence of the two samples makes the remaining cross component orthogonal to these terms, with variance $4\beta^2 \operatorname{tr}(\Sigma_A \Sigma_B)/(n_v N_v)$. Adding these mutually orthogonal contributions proves the exact variance formula in Eq. (47). □

D Supplementary Learning-Theoretic Guarantees

The main paper gives a sharp local fixed-dimensional analysis. This section records two additional consistency results not used in Theorem 3.7: a global fixed-power bound and a localized nonparametric extension. They establish complementary learnability guarantees rather than the post-conformal efficiency comparisons developed above.

D.1 A distribution-free global excess-risk bound

Let ρ_τ be $L_\rho \le 1$ Lipschitz in its prediction argument. Assume $|q(X)| \le B_q$, $|S| \le B_S$, and $|S^\dagger|, |\widetilde{S}| \le B_\dagger$. Set $B_0 = B_q + B_S$, $B_1 = B_q + B_\dagger$, and let $q^\star_{\mathcal{Q}} \in \arg\min_{q\in\mathcal{Q}} \mathcal{R}(q)$.

Theorem D.1 (Uniform PPI-ERM bound)**.** *For fixed λ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathcal{R}(\widehat{q}_\lambda) - \mathcal{R}(q^\star_{\mathcal{Q}}) \le 4(1+|\lambda|)L_\rho \mathfrak{R}_n(\mathcal{Q}) + 4|\lambda| L_\rho \mathfrak{R}_N(\mathcal{Q}) + 2(B_0 + |\lambda| B_1)\sqrt{\frac{\log(4/\delta)}{2n}} + 2|\lambda| B_1 \sqrt{\frac{\log(4/\delta)}{2N}}. \tag{59}$$

Proof. The deviation from the true risk is the sum of a paired empirical process and an independent pool process. The ERM reduction, symmetrization, contraction, bounded-difference concentration, and a union bound give Eq. (59). □

This bound establishes consistency rather than dominance; the sharp local improvement criterion is given by Corollaries B.3 and 3.8.

D.2 A radius-dependent nonparametric extension

Let $q^\star \in \mathcal{Q}$ and $\mathcal{E}(q) = \mathcal{R}(q) - \mathcal{R}(q^\star)$. For $\lambda = 1$, define

$$\Delta_q = (\ell_q - \ell_{q^\star})(X,S) - (\ell_q - \ell_{q^\star})(X,S^\dagger), \qquad \widetilde{g}_q = (\ell_q - \ell_{q^\star})(\widetilde{X},\widetilde{S}).$$

Knight's identity gives

$$\Delta_q = \int_{q^\star(X)}^{q(X)} \{\mathbb{1}(S \le u) - \mathbb{1}(S^\dagger \le u)\}\, du.$$

Let D_q indicate disagreement somewhere on this segment, so $|\Delta_q| \le |q - q^\star| D_q$. Assume $|q - q^\star| \le B_Q$ and local curvature $\mathcal{E}(q) \ge c_0 \|q - q^\star\|_2^2$. Define

$$\eta(r) = \sup_{\substack{q:\mathcal{E}(q)\le r \\ q\ne q^\star}} \frac{\mathbb{E}[(q-q^\star)^2 D_q]}{\mathbb{E}[(q-q^\star)^2]}. \tag{60}$$

Assumption D.2 (Localized VC-type deviations)**.** There are constants $A \ge e, D \ge 1, C_0$ such that, simultaneously for relevant r and with probability at least $1-\delta$,

$$\sup_{\mathcal{E}(q)\le r} |(P_n - P)\Delta_q| \le C_0 \left[\sqrt{\frac{\eta(r) r V_n}{c_0 n}} + \frac{B_Q V_n}{n}\right], \tag{61}$$

$$\sup_{\mathcal{E}(q)\le r} |(\widetilde{P}_N - P^\dagger)\widetilde{g}_q| \le C_0 \left[\sqrt{\frac{r V_N}{c_0 N}} + \frac{B_Q V_N}{N}\right],$$

where $V_n = D\log(An) + \log(4/\delta)$ and $V_N = D\log(AN) + \log(4/\delta)$.

Theorem D.3 (Radius-dependent fixed point)**.** *If $r_\star$ satisfies, for every $r \ge r_\star$,*

$$C_0 \left[\sqrt{\frac{\eta(r) r V_n}{c_0 n}} + \sqrt{\frac{r V_N}{c_0 N}} + \frac{B_Q V_n}{n} + \frac{B_Q V_N}{N}\right] \le r/4, \tag{62}$$

then with probability at least $1-\delta$, $\mathcal{E}(\widehat{q}_1) \le r_\star$. If $\eta(r) \le \bar{\eta}$ locally, one may take

$$r_\star \lesssim \frac{\bar{\eta} V_n}{c_0 n} + \frac{V_N}{c_0 N} + \frac{B_Q V_n}{n} + \frac{B_Q V_N}{N}. \tag{63}$$

Proof. The ERM basic inequality bounds $\mathcal{E}(\widehat{q}_1)$ by the two empirical deviations. Applying Assumption D.2 on dyadic excess-risk shells excludes every shell above $r_\star$ via Eq. (62); solving the resulting inequalities gives Eq. (63). □

Corollary D.4 (A local bridge from nonparametric risk to coverage)**.** *Under the smoothness conditions of Corollary 3.5 and local curvature, suppose additionally that $\|\widehat{q}_1 - q^\star\|_\infty \le \varepsilon_0$ on the event of interest, where ε_0 is inside the proposition's local neighborhood. Then*

$$\|\mathcal{P}(\widehat{q}_1, \tau) - \tau\|_2^2 \le C\,\mathcal{E}(\widehat{q}_1). \tag{64}$$

Consequently, if the required sup-norm localization also holds with high probability, the population-conformalized conditional error is $O_{\mathrm{Pr}}(r_\star)$. Independent finite calibration is governed separately by Proposition 3.2 and Proposition C.1; it is not absorbed into the learning-risk rate.

Proof. For $e = q - q^\star$, Corollary 3.5 with $v = 0$ gives $\|\mathcal{P}(q, \tau) - \tau\|_2 \le C\|e\|_2 + C\|e\|_4^2$. On the local sup-norm event, $\|e\|_4^2 \le \varepsilon_0 \|e\|_2$; after squaring, local curvature converts the right-hand side to $C\mathcal{E}(q)$. □