WHO WINS WHERE? CONFORMAL MODEL COMPARISON FOR LOCAL SUPERIORITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard model comparison is global, aggregating losses across the covariate space to declare a single winner. This can obscure heterogeneous performance, where different models are preferable in different regions. We introduce *conformalized local model comparison*, a split-sample framework for constructing calibrated local winner maps. Under exchangeability, our goal is to provide finite-sample, distribution-free control of the marginal false winner rate while allowing declarations to depend on covariates; the guarantee is marginal because pointwise conditional control is unattainable in this setting. Given a model comparison score, such as the difference between two squared losses, the method estimates local centers and scales from out-of-sample scores and conformally calibrates the remaining uncertainty. At a test point, the procedure declares a local winner only when a one-sided conformal upper bound falls below zero. We instantiate the framework with a pretrained TabPFN localizer whose regression head is fine-tuned to maximize conformal gain. For any choice of localizer, we prove that the error control applies to one-sided false declarations on the realized future comparison score. We further give sufficient conditions for pointwise recovery of the expected winner away from tie boundaries, show that aggregate comparison can sharply contradict the prevalence of local superiority, and derive a squared-loss bias–variance decomposition of local wins. Synthetic and real-world experiments show that the method recovers heterogeneous winner regions and abstains under uncertainty, identifying local gains that global comparison can miss. We provide our code and data in the supplementary material.

1 INTRODUCTION

Model comparison is central to machine learning. Standard practice averages the losses of two trained predictors over a validation set and declares a global winner. This simple workflow is often appropriate when one model dominates uniformly, but not all practically important comparisons are global. A structured or theory-guided model may perform better where its inductive bias aligns with the data-generating mechanism, while a flexible black-box model may perform better elsewhere. Likewise, expert routing may favor different predictors in different covariate subspaces.

Local model comparison asks which model is superior near a target covariate x_0, rather than which model wins on average. The local expected winner is defined by the sign of the conditional mean score $\mu(x) = \mathbb{E}[S(X, Y) \mid X = x]$, such as the difference between two squared losses, where negative values favor model A and positive values favor model B. Majority and quantile winners define the local winner beyond the mean. A general localizer estimates these targets from data.

This paper studies *conformalized local model comparison*. A local comparison is more sensitive to heterogeneity than a global one, but also more variable, so we calibrate it conformally. A winner is declared only when a one-sided conformal upper bound falls below zero. This construction is deliberately modest: it is a prediction bound for the *realized future score* with finite-sample *marginal* protection against errors, not a confidence interval for $\mu(x_0)$ itself. Unlike tests of conditional predictive ability, which provide asymptotic confidence statements about the expected loss difference under assumptions on the data-generating process (Giacomini & White, 2006), this bound requires only exchangeability, so its control is marginal over the covariates (Foygel Barber et al., 2021).

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.

The local center and scale can come from any localizer fitted on the estimation split. An adaptive localizer can maintain power when irrelevant coordinates weaken geometric neighborhoods. Our main instantiation uses TabPFN-3.5 (Hollmann et al., 2025; Jäger et al., 2026) with a frozen backbone and a regression head fine-tuned to maximize conformal gain.

Our theory has three parts. First, a neighborhood-average estimator recovers the sign of $\mu(x_0)$ asymptotically away from tie boundaries. Second, aggregate comparison targets the global mean score and can sharply contradict the prevalence of local superiority. Third, under squared loss, the local mean score decomposes into differences in squared bias and variance, making the framework natural for structured-versus-flexible comparisons.

In summary, our contributions are: (i) a local model-comparison formulation based on the conditional mean score $\mu(x)$ with general localizers, together with secondary majority and quantile descriptors; (ii) a split-sample, locally centered conformal comparison rule with finite-sample marginal control for false local winner declarations on the realized future score; (iii) a TabPFN localizer with a frozen backbone and a gain-trained regression head tailored to identifying regions of local model superiority; and (iv) theoretical results on consistency, global–local mismatch, and a bias–variance decomposition that explain when and why winner regions emerge.

2 Related Literature

Conformal prediction, localization, and weighting. Conformal prediction provides finite-sample, distribution-free predictive guarantees under exchangeability (Vovk et al., 2005; Shafer & Vovk, 2008; Lei et al., 2018; Angelopoulos & Bates, 2023). Methods adapt to heterogeneity through conformalized quantile regression (Romano et al., 2019), weighted conformal prediction under covariate shift (Tibshirani et al., 2019), and localized conformal prediction (Guan, 2023; Hore & Barber, 2025). Another line incorporates conformal objectives into training to improve calibrated-output efficiency (Colombo & Vovk, 2020; Stutz et al., 2021; Einbinder et al., 2022; Bai et al., 2022). Our paper draws on both ideas, adapting calibration to heterogeneity and training for the calibrated output, but the inferential object is different: we calibrate a *comparison score between two models*, not a prediction interval for a single model, and use a gain-based training objective to encourage useful local winner declarations.

Localized model selection and post-selection validity. Recent conformal work addresses model selection. Localized conformal model selection (Wang & Wang, 2026) provides a framework for choosing models that yield efficient conformal intervals while preserving coverage, and Liang et al. (2026) study predictive validity following data-dependent model choices. Our work shares their view that model suitability is heterogeneous, but our objective is complementary. Rather than optimizing interval length or characterizing post-selection predictive coverage, we calibrate a comparison score to control erroneous declarations on its realized future sign (Jin & Candès, 2023).

Model comparison and routing. The broad motivation also connects to conditional predictive ability testing and local model routing. Mixture-of-experts and algorithm-selection perspectives emphasize that different models may be preferable in different regions of the input space (Jordan & Jacobs, 1994). Conditional predictive-ability tests examine loss differences using conditioning information (Giacomini & White, 2006). These tests give a confidence statement about the expected loss difference given the conditioning information, and their validity is asymptotic under mixing and moment assumptions. Our rule instead bounds the realized score at a new point with a locally fitted center and scale, and this prediction bound is valid in finite samples under exchangeability alone, as in split conformal prediction. This control is marginal over X, because no informative distribution-free bound can hold at each fixed covariate value (Foygel Barber et al., 2021).

Localizers and learning on tabular data. The local structure of the model-comparison score can be learned by nearest-neighbor methods, or tree ensembles such as random forests and their extensions to heterogeneous treatment effects (Breiman, 2001; Wager & Athey, 2018; Athey et al., 2019), or tabular neural networks. TabNet selects features through attention masks (Arik & Pfister, 2021) and RealMLP weights them through learnable input scaling (Holzmüller et al., 2024). More recently, pretrained tabular foundation models such as TabPFN exploit transfer across tasks to generalize to unseen tabular problems (Hollmann et al., 2025; Jäger et al., 2026). TabArena benchmarks these

model families across diverse tabular tasks (Erickson et al., 2026). We use TabPFN as our main localizer and adapt its regression head with a gain-based objective for local model comparison.

3 Problem Setup and Local Model Comparison

Let $(X, Y) \in \mathcal{X} \times \mathbb{R}$ be a random covariate–response pair. We compare two fitted models, A and B, with predictors $\hat{f}_A, \hat{f}_B : \mathcal{X} \to \mathbb{R}$. For each model $m \in \{A, B\}$, let $L_m(X, Y)$ denote its loss at (X, Y). We use the generic score $S(X, Y) := s\big(L_A(X, Y), L_B(X, Y); X\big)$, where negative scores favor A, positive scores favor B, and zero denotes a tie. We assume $s(\ell_A, \ell_B; x)$ is weakly increasing in ℓ_A and weakly decreasing in ℓ_B. Examples include the gap $s_{\text{gap}}(\ell_A, \ell_B) = \ell_A - \ell_B$, the log-ratio $s_{\text{ratio}}(\ell_A, \ell_B) = \log((\ell_A + \kappa)/(\ell_B + \kappa))$ with $\kappa > 0$, and the standardized gap $s_{\text{std}}(\ell_A, \ell_B; x) = (\ell_A - \ell_B)/v(x)$ with $v(x) > 0$.

Local expected winner. The local expected winner is defined through the conditional mean score $\mu(x) := \mathbb{E}[S(X, Y) \mid X = x]$.

Definition 1 (Local expected winner). *Model A is a* local expected winner *at x_0 if $\mu(x_0) < 0$. The local winner map is $\mathcal{W}(x) = A\,\mathbb{I}_{\{\mu(x)<0\}} + B\,\mathbb{I}_{\{\mu(x)>0\}} + \text{tie} \cdot \mathbb{I}_{\{\mu(x)=0\}}$.*

Local majority and quantile winners. Two further notions describe local agreement beyond the mean. Let $F_x(t) := \mathbb{P}(S(X, Y) \le t \mid X = x)$, and define $\pi_A(x) := \mathbb{P}(S(X, Y) < 0 \mid X = x) = F_x(0^-)$ and $Q_{1-\beta}(x) := \inf\{t \in \mathbb{R} : F_x(t) \ge 1 - \beta\}$ for $\beta \in (0, 1)$. Model A is the local majority winner when $\pi_A(x) > 1/2$, and the local quantile winner when $Q_{1-\beta}(x) < 0$.

General localizers. A localizer estimates these targets from data. It is fitted to the out-of-sample scores $\{(X_j, S_j) : j \in I_{\text{est}}\}$ of the estimation split defined in Section 4 and returns a local center $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$ and a positive scale $\widehat{s} : \mathcal{X} \to (0, \infty)$. The estimated expected-winner map replaces μ with $\widehat{\mu}$. A distributional localizer also returns a fitted distribution $\widehat{F}_x$, from which all three winner maps are estimated. The targets themselves do not depend on the localizer.

4 Conformalized Local Model Comparison

4.1 Conformal calibration and declaration

We present the rule for two models A and B. For $M > 2$ models, Appendix A.5 calibrates pairwise comparisons jointly, which returns a partial order of the models. To attach finite-sample inference to $S(X, Y)$, split an i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^n$ into three disjoint parts: I_{fit} only trains the competing predictors $\hat{f}_A, \hat{f}_B$; I_{est} fits and tunes the localizer from *out-of-sample* scores; and I_{cal} is used only for conformal calibration. This avoids using the same observations to fit the models and estimate their local scores from in-sample losses. Conditional on the models fitted on I_{fit}, the population target $\mu(x)$ and out-of-sample scores S_j for $j \in I_{\text{est}}$ are evaluated as defined in Section 3.

On the calibration split I_{cal} of size n_{cal}, we compute the scores S_i analogously and form the localized residuals $R_i := (S_i - \widehat{\mu}(X_i))/\widehat{s}(X_i)$, with the center and scale of Section 3. Setting $\widehat{s} \equiv 1$ gives the unstandardized rule; the fitted scale gives the standardized rule. Let $R_{(1)} \le \cdots \le R_{(n_{\text{cal}})}$ be the sorted residuals. Setting $k_\alpha := \lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil$, the conformal threshold is $\widehat{q}_{1-\alpha} := R_{(k_\alpha)}$, with $\widehat{q}_{1-\alpha} := \infty$ if $k_\alpha > n_{\text{cal}}$. For a test covariate $x_0 \in \mathcal{X}$, the conformal upper bound is $\widehat{U}_\alpha(x_0) := \widehat{\mu}(x_0) + \widehat{s}(x_0)\widehat{q}_{1-\alpha}$. We declare model A the local winner, written $\hat{w}_\alpha(x_0) = A$, when $\widehat{U}_\alpha(x_0) < 0$. This one-sided rule controls false declarations in favor of A; the analogous declaration for model B follows by interchanging A and B.

Proposition 1 (Finite-sample marginal control for the realized score). *Let (X_{n+1}, Y_{n+1}) be a new test point. Suppose the calibration observations and the test point are exchangeable conditional on the training and estimation splits, which determine the fitted predictors and the measurable center $\widehat{\mu}$ and scale $\widehat{s}$, with $\widehat{s}(x) > 0$. Then the realized score $S_{n+1} := S(X_{n+1}, Y_{n+1})$ satisfies*

$$\mathbb{P}\Big\{S_{n+1} \le \widehat{U}_\alpha(X_{n+1})\Big\} \ge 1 - \alpha, \qquad \mathbb{P}\Big\{\widehat{U}_\alpha(X_{n+1}) < 0,\ S_{n+1} \ge 0\Big\} \le \alpha.$$

Proposition 1 deliberately concerns the *realized future score*, not a finite-sample confidence interval or hypothesis test for $\mu(x_0)$. The one-sided rule's false winner rate for this realized score is at most α, marginally over calibration and a new test point. Because it relies only on residual exchangeability, the guarantee holds for any comparison score and center–scale localizer.

4.2 Gain-guided fine-tuning of the TabPFN head

Our main localizer is TabPFN-3.5 (Hollmann et al., 2025; Jäger et al., 2026), a pretrained tabular foundation model that generalizes to unseen tabular tasks through in-context learning. However, its pretraining targets predictive accuracy and does not directly optimize the value of the declared region. We therefore freeze the TabPFN backbone and update only its native regression head, the regression MLP and its output projection, to favor declarations with positive realized total gain.

Write $\bar{S}_{\mathrm{est}}$ and $\mathrm{sd}(S_{\mathrm{est}}) > 0$ for the mean and standard deviation of the estimation scores. Training uses $\widetilde{S} = (S - \bar{S}_{\mathrm{est}})/\mathrm{sd}(S_{\mathrm{est}})$ with declaration threshold $b = -\bar{S}_{\mathrm{est}}/\mathrm{sd}(S_{\mathrm{est}})$. An 80% training portion of I_{est} forms five folds, each embedded as queries against the others; the other 20% is a stopping set.

For an inner episode, divide a fold's queries into an inner calibration set C and an evaluation set V. Given standardized head outputs $\widetilde{\mu}_\phi, \widetilde{s}_\phi$, compute the same finite-sample residual order statistic on C and let $\widetilde{U}_{\phi,\alpha}(X_i) = \widetilde{\mu}_\phi(X_i) + \widetilde{s}_\phi(X_i)\widetilde{q}_{C,1-\alpha}$. The episode objective is

$$\mathcal{L}_{\mathrm{gain}}(\phi) = \frac{1}{|V|}\sum_{i\in V}(\widetilde{S}_i - b)\,\mathrm{sigmoid}\left(\frac{b - \widetilde{U}_{\phi,\alpha}(X_i)}{T}\right). \tag{1}$$

Since $\widetilde{S}_i - b = S_i/\mathrm{sd}(S_{\mathrm{est}})$, minimizing this objective maximizes a soft version of the realized total gain. The sigmoid with temperature $T = 0.1$ smooths the declaration indicator, and gradients pass through the empirical quantile, via the residual it selects, to the head outputs. The episodes use only the estimation data, not the outer calibration split. To preserve the predictive moments, we constrain two empirical moment risks on an index set I,

$$\mathcal{R}_{\mathrm{MSE}} = \frac{1}{|I|}\sum_{i\in I}\big(\widetilde{S}_i - \widetilde{\mu}_\phi(X_i)\big)^2, \qquad \mathcal{R}_{\mathrm{G}} = \frac{1}{2|I|}\sum_{i\in I}\left[\frac{(\widetilde{S}_i - \widetilde{\mu}_\phi(X_i))^2}{\widetilde{s}_\phi(X_i)^2} + 2\log\widetilde{s}_\phi(X_i)\right].$$

Here $\mathcal{R}_{\mathrm{MSE}}$ is the mean squared error of the center, and $\mathcal{R}_{\mathrm{G}}$ is the Gaussian negative log-likelihood of the center and scale. For $k \in \{\mathrm{MSE}, \mathrm{G}\}$, let $\mathcal{R}_{0,k}$ be the pretrained head's risk and let $c_k = \mathcal{R}_{0,k} + 0.05|\mathcal{R}_{0,k}|$ be its cap. The head minimizes the Lagrangian $\mathcal{L}(\phi) = \mathcal{L}_{\mathrm{gain}}(\phi) + \sum_k \lambda_k\{\mathcal{R}_k(\phi) - c_k\}$, where each multiplier $\lambda_k \geq 0$ is updated by dual ascent with step size η, $\lambda_k \leftarrow \max\{0, \lambda_k + \eta(\mathcal{R}_k - c_k)\}$, so that it grows only while the training risk exceeds its cap.

5 Theoretical Properties

We study consistency for a neighborhood-average localizer, a tractable instance of our framework. The first two results condition on the independently fitted predictors, keeping evaluation observations i.i.d. and $\mu(x)$ well defined. Proposition 3 reintroduces training randomness for an ex-ante version of the expected local score under squared loss. Let $K_r : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ denote neighborhood weights with bandwidth $r > 0$. For this localizer, the center $\widehat{\mu}$ of Section 4 is the weighted score average (2) computed from I_{est}.

Asymptotic correctness for local expected superiority. We first study whether a localized estimator recovers the correct sign of $\mu(x_0)$ at a target covariate x_0. Given an *independent evaluation sample* and a bandwidth sequence $r_n \to 0$, the neighborhood-average estimator is

$$\hat{\mu}_n(x_0) := \frac{\sum_{i=1}^n K_{r_n}(X_i, x_0)S_i}{\sum_{i=1}^n K_{r_n}(X_i, x_0)}, \tag{2}$$

where $S_i = S(X_i, Y_i)$ is the realized score. Let the true local winner be $\delta^\star(x_0) := A$ if $\mu(x_0) < 0$ and B if $\mu(x_0) > 0$. Its empirical counterpart $\hat{\delta}_n(x_0)$ equals A if $\hat{\mu}_n(x_0) < 0$ and B otherwise; this tie convention differs from $\mathcal{W}$ only on events of measure zero for continuous scores. The sign is not identifiable on the tie boundary $\mu(x_0) = 0$, which the theorem excludes.

Theorem 1 (Asymptotic correctness of local winner estimation)**.** *Fix an interior point $x_0 \in \mathcal{X}$. Assume: (1) $X \in \mathbb{R}^d$ has a density p_X that is continuous at x_0 with $p_X(x_0) > 0$; (2) the localizer is $K_r(x, x_0) = r^{-d}K((x - x_0)/r)$, where K is bounded, nonnegative, compactly supported, and satisfies $\int_{\mathbb{R}^d} K(u)\,du > 0$; (3) $\mu(x)$ is continuous at x_0; (4) there exists a neighborhood $U(x_0)$ of x_0 such that $\sup_{x\in U(x_0)} \mathbb{E}[S(X,Y)^2 \mid X = x] < \infty$; and (5) $r_n \to 0$ and $nr_n^d \to \infty$. Then $\hat{\mu}_n(x_0) \xrightarrow{p} \mu(x_0)$. Consequently, if $\mu(x_0) \neq 0$, then $\mathbb{P}(\hat{\delta}_n(x_0) = \delta^\star(x_0)) \longrightarrow 1$. The same applies to the estimation-split version of (2) with n replaced by $|I_{\text{est}}|$.*

Aggregate comparison targets the global mean, not local prevalence. Define the global mean score $\mu_{\text{glob}} := \mathbb{E}[S(X,Y)]$ and the aggregate empirical score $\bar{S}_n := \frac{1}{n}\sum_{i=1}^n S_i$. The aggregate selector chooses $\hat{\delta}_n^{\text{agg}} := A$ if $\bar{S}_n < 0$ and B otherwise. The prevalence of local superiority for model A is $\rho_A := \mathbb{P}(\mu(X) < 0)$, which depends only on the sign of μ.

Proposition 2 (Aggregate selection can disagree with local superiority prevalence)**.** *Suppose $\mathbb{E}[|S(X,Y)|] < \infty$. Let the true aggregate winner be $\delta^{\text{agg},\star} := A$ if $\mu_{\text{glob}} < 0$ and B if $\mu_{\text{glob}} > 0$. If $\mu_{\text{glob}} \neq 0$, then $\mathbb{P}(\hat{\delta}_n^{\text{agg}} = \delta^{\text{agg},\star}) \longrightarrow 1$. Moreover, for any $q \in (1/2, 1)$, there exists a joint distribution of (X, S) such that $\mathbb{P}(\mu(X) < 0) \geq q$ but $\mathbb{E}[S] > 0$. In such cases, $\mathbb{P}(\hat{\delta}_n^{\text{agg}} = B) \longrightarrow 1$ even though model A is locally superior on at least a q-fraction of the covariate distribution.*

Proposition 2 makes an elementary point that standard model evaluation often overlooks. The global winner is not determined by the proportion of local winners. Aggregate averaging weighs the size of the scores, so large losses in a small covariate region can dominate the sum. Local prevalence depends only on the share of the covariate distribution where $\mu(x) < 0$. Global selection can therefore reject a model that is the better choice for 99% of the population.

Ex-ante bias–variance interpretation for squared loss. We now reintroduce training randomness to understand *why* regions of superiority emerge, using an ex-ante expected local score over both the training procedure and the test distribution.

Proposition 3 (Squared-loss local bias–variance decomposition)**.** *Assume $Y = f^\star(X) + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] = 0$, and assume the training sample $\mathcal{D}$ used to fit the predictors is independent of the fresh test pair (X, Y). Let the training-random squared-loss gap be $S_{\mathcal{D}}(X,Y) := (\hat{f}_A(X) - Y)^2 - (\hat{f}_B(X) - Y)^2$. For $m \in \{A, B\}$, define $\text{Bias}_m(x) := \mathbb{E}_{\mathcal{D}}[\hat{f}_m(x)] - f^\star(x)$ and $\text{Var}_m(x) := \text{Var}_{\mathcal{D}}(\hat{f}_m(x))$. If $\mathbb{E}_{\mathcal{D}}[\hat{f}_m(x)^2] < \infty$ and $\mathbb{E}[\varepsilon^2 \mid X = x] < \infty$, then the ex-ante conditional mean score $\mu_{\text{ex}}(x) := \mathbb{E}_{\mathcal{D},Y|X=x}[S_{\mathcal{D}}(X,Y) \mid X = x]$ satisfies*

$$\mu_{\text{ex}}(x) = \text{Bias}_A^2(x) - \text{Bias}_B^2(x) + \text{Var}_A(x) - \text{Var}_B(x). \tag{3}$$

Proposition 3 explains where local wins come from. Because irreducible noise cancels, the ex-ante local mean score reduces to squared-bias and variance differences between a structured model with high bias and low variance (e.g., linear regression) and a flexible model with low bias and high variance (e.g., gradient boosting). Where the data-generating mechanism aligns with the structured model's inductive bias, its variance advantage makes it the local winner. In nonlinear regions, the flexible model's bias reduction outweighs its variance, and the local winner changes.

6 Experiments

In this section, synthetic experiments test recovery of heterogeneous winner regions (Theorem 1, Proposition 2), declaration reliability (Proposition 1), the bias–variance mechanism (Proposition 3), and localizer choice as dimension grows. Real regression and classification experiments (Section 6.2) test whether conformal declarations select regions with positive realized gain. Appendix B.6 examines the multiple-model extension of Appendix A.5.

6.1 Synthetic data experiments

We use five one-dimensional synthetic designs, in which winner regions can be inspected visually and compared with an oracle. In all designs, $X \sim \text{Unif}[0,1]$ and $Y = f^\star(X) + \varepsilon$. Model A is a

gradient-boosted tree (LightGBM) and model B is ordinary linear regression. Each replicate has $n = 5000$ observations, split evenly into fitting, estimation, calibration, and test splits. Results are averaged over five random seeds, with $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$ and $\alpha = 0.10$ for the tables. The primary score is the squared-loss gap, where $S < 0$ favors the boosted tree. We evaluate local winners with the post-fit oracle $\mu(x) = (\hat{f}_A(x) - f^\star(x))^2 - (\hat{f}_B(x) - f^\star(x))^2$, comparing the already-fitted predictors at each covariate value. Full data-generating process (DGP) formulas, design motivations, and model settings are in Appendix B.1, and evaluation metrics in Appendix B.2.

The local conformal methods use TabPFN-gain (Section 4.2) for the center and scale, and the unstandardized and standardized rules share its center. The three non-conformal winner maps are read from the TabPFN predictive distribution. Implementation details of all localizers are in Appendix C.

Results #1: Local comparison recovers winner regions hidden by global averages. Figure 1 illustrates global–local mismatch in Case 3 (Table 2). The data-generating function is linear outside $[0.42, 0.58]$ but oscillates strongly inside. Model A has its main advantage inside this island; B is preferable over much of the remaining space. Global mean assigns one winner everywhere from the aggregate mean score. The island's large loss differences make it declare A everywhere, missing most of the B-favorable region. We evaluate the three non-conformal local winner maps of Section 3: the expected, majority, and quantile winners. In Case 3, Global mean attains only 0.299 winner accuracy, compared with 0.956 for the local expected winner and 0.938 for the quantile winner; full results are in Table 4. In Figure 1, the expected winner follows the oracle $\mu(x)$ sign, the majority winner is more sensitive to local score variation, and the quantile winner is more conservative. These local rules reveal winner regions that are invisible to the global mean, as Proposition 2 predicts.

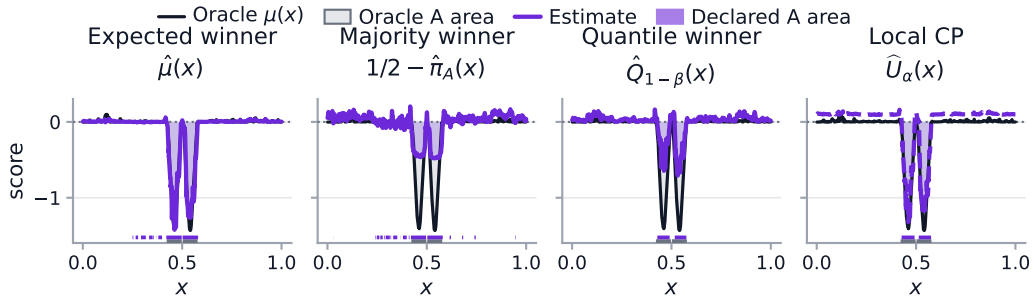

Figure 1: Case 3. Panels 1–3: non-CP expected, majority, and quantile winner ($\beta = 0.20$), from the TabPFN predictive distribution. Panel 4: Local CP, declaring A where $\widehat{U}_\alpha$ (dashed) is negative.

Results #2: Conformal calibration controls false winner declarations. Figure 2 (top) plots false winner rate against nominal α for all five cases. The conformal methods follow the one-sided upper-bound rule in Section 4, declaring A only when $\widehat{U}_\alpha(x) < 0$. Global CP uses a constant center and unit scale; Local CP uses the fitted center $\widehat{\mu}(x)$ with unit scale; and Local scaled CP uses both $\widehat{\mu}(x)$ and the fitted scale $\widehat{s}(x)$. Mean conformal false winner rates stay below the diagonal in all five cases, consistent with Proposition 1. In contrast, non-conformal winner maps can over-declare when realized scores are noisy.

Figure 1's first and last panels distinguish identifying from certifying a local winner. The non-conformal rule selects A wherever the estimated mean is negative, including short intervals left of the island where A is not the oracle winner. The conformal rule requires its upper bound to be negative, conservatively selecting inside the island.

Local centering and scaling generally improve conformal declaration efficiency. Using one center over the covariate space, Global CP declares A everywhere or abstains everywhere. Local methods instead select A-favorable regions. Figure 2's lower rows report selection rate and oracle A power. Error control leads Local CP to select fewer points than the oracle A region, but more spatially targeted points than Global CP. In the heteroscedastic Case 5, local scaling substantially improves selection and power over the unstandardized rule by preventing high-variance regions from dominating the calibration quantile. The unstandardized rule is better in Case 2, where the local scale absorbs the subtle score signal.

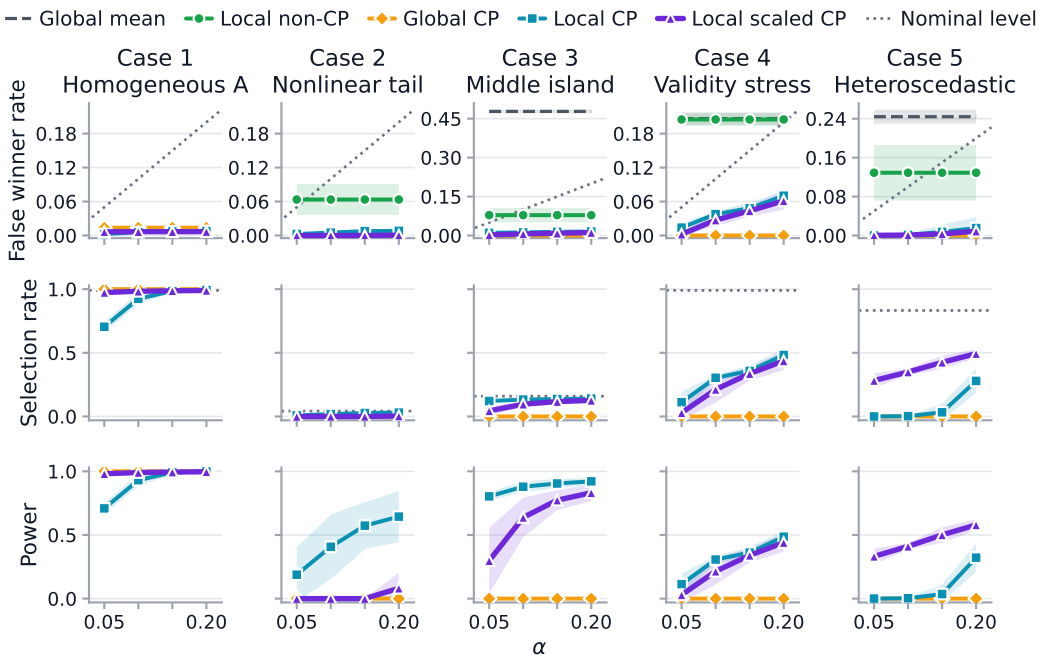

Figure 2: False winner rate (top; diagonal: target), selection rate (middle; dotted: oracle A-winning rate), and oracle A power (bottom). Mean and standard deviation over five seeds.

Results #3: Diagnostics for the theoretical mechanisms. Figure 3 shows the bias–variance decomposition in Case 3. Outside the island, the linear model has low variance and small bias, so B is preferable; inside the island, the boosted tree's bias reduction dominates its variance cost, so A wins. The decomposition and the Monte Carlo mean gap agree, as Proposition 3 predicts.

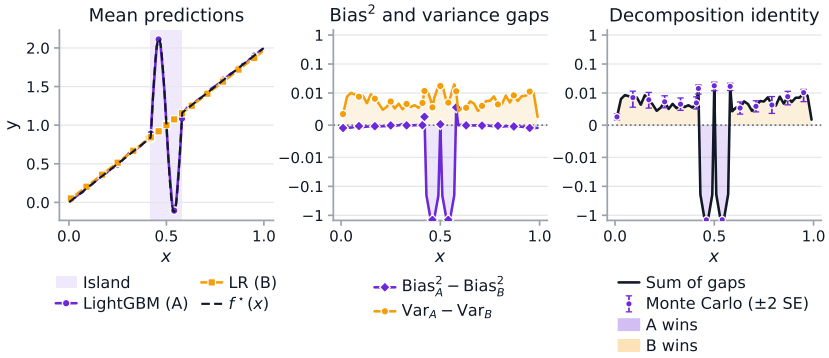

Figure 3: Bias–variance decomposition on Case 3. Left: signal and mean predictions. Middle: squared-bias and variance differences; negative values favor A. Right: their sum and the Monte Carlo mean loss gap, ± 2 standard errors over 100 fits. Middle and right: symmetric-log scale.

Results #4: Different localizer behaviors as the covariate dimension grows. To compare localizers, we append independent nuisance coordinates and repeat all five cases at $d \in \{1, 10, 30, 100\}$. We compare nearest neighbors, random forests, LightGBM, MLP, TabNet, pretrained TabPFN, and TabPFN-gain under the same standardized conformal rule. Table 1 reports five-case average power. As nuisance dimensions grow, nearest neighbors and the MLP lose much of their power, while random forests and TabPFN retain substantially more. At $d = 100$, their average powers are 0.415 and 0.394, against 0.196 for nearest neighbors, and LightGBM falls in between. Adaptive tree-based and pretrained localizers thus handle nuisance dimensions better than fixed geometric neighborhoods. TabPFN-gain achieves the highest average power at $d = 10, 30, 100$, is close to the best at $d = 1$,

and improves on pretrained TabPFN at all four dimensions, from 0.394 to 0.428 at $d = 100$. Its average power stays stable from $d = 1$ to 100 (0.451 and 0.428), as do the case-level results in Appendix B.5. A pretrained localizer thus preserves informative local comparison in high dimensions, while gain-guided fine-tuning improves declaration efficiency.

Table 1: Standardized conformal power by nuisance dimension at $\alpha = 0.10$, averaging five cases within seed. Mean and standard deviation over five seeds; bold: column maximum.

Case	Family	$d = 1$	$d = 10$	$d = 30$	$d = 100$
Average (5 cases)	k-NN	0.418 ± 0.013	0.216 ± 0.010	0.200 ± 0.011	0.196 ± 0.003
	Forest	0.428 ± 0.032	0.424 ± 0.028	0.433 ± 0.024	0.415 ± 0.021
	LightGBM	0.443 ± 0.032	0.311 ± 0.036	0.285 ± 0.050	0.325 ± 0.017
	MLP	$\mathbf{0.456 \pm 0.009}$	0.215 ± 0.025	0.217 ± 0.030	0.201 ± 0.005
	TabNet	0.270 ± 0.058	0.258 ± 0.028	0.243 ± 0.032	0.221 ± 0.038
	TabPFN	0.404 ± 0.052	0.385 ± 0.021	0.434 ± 0.012	0.394 ± 0.044
	TabPFN-gain	0.451 ± 0.038	$\mathbf{0.425 \pm 0.022}$	$\mathbf{0.469 \pm 0.032}$	$\mathbf{0.428 \pm 0.027}$

6.2 Real data experiments

We evaluate four public regression benchmarks: Concrete, Auto MPG, BlogFeedback, and Facebook. Model A is a one-hidden-layer MLP, with architecture selected per dataset by held-out fitting-split validation loss; B is linear regression. We seek regions where linear regression beats the tuned MLP, hidden by the global ranking. Dataset statistics are in Table 5. Since no oracle map is available, we report model B's selection rate and conditional gain $\mathbb{E}[L_A - L_B \mid \hat{w}_\alpha = B]$. Splits and localizers follow the synthetic setup.

Results #5: Model A vs. model B winning regions. Figure 4 shows declarations in a PCA projection; intermediate colors reflect variation across seeds. Global mean assigns one model everywhere within each seed. Local non-conformal maps reveal subregions hidden by the global decision. Columns 3–4 highlight our method's advantage: Global CP abstains, while Local CP identifies B-favorable regions. Compared to the non-conformal baseline, our conformalized approach yields smaller, statistically controlled subregions.

Results #6: Selection vs. conditional gain. Figure 5 plots conditional gain against selection rate across α. Increasing α expands declarations, typically reducing their concentration in high-gain regions. Non-conformal methods select more points but achieve lower conditional gain, reflecting broad declarations without false-winner control. Global CP abstains entirely. Local CP achieves the highest conditional gain, exceeding its non-conformal counterpart. This confirms that its abstention mechanism isolates a smaller region with a larger realized advantage for B.

Results #7: Classification. We also test binary classification on Mozilla4, Page-blocks, and GiveMeSomeCredit. Model A is an MLP classifier tuned by the Brier score, model B is logistic regression, and $L_m = (\hat{p}_m(X) - Y)^2$ is the Brier loss of the positive-class probability. Figure 5 (bottom) shows the same pattern as regression: average losses favor the MLP, but Local CP identifies smaller regions with positive mean gain for the linear model. Thus, extending to classification requires changing only the comparison-score loss, demonstrating the generality of our framework.

7 Conclusion

We presented a conformal framework for local model comparison based on the conditional mean score and conformal calibration. We used TabPFN as the localizer and fine-tuned its regression head for the conformal gain. The resulting localizer remains stable as nuisance dimensions grow and improves declaration power over pretrained TabPFN. The theory establishes asymptotic consistency for neighborhood-based winner estimation, the disconnect between aggregate comparison and the prevalence of local superiority, and an ex-ante bias–variance decomposition that explains local advantages. On synthetic and real regression and classification data, local winner maps with conformal declarations reveal heterogeneous performance that global metrics miss.

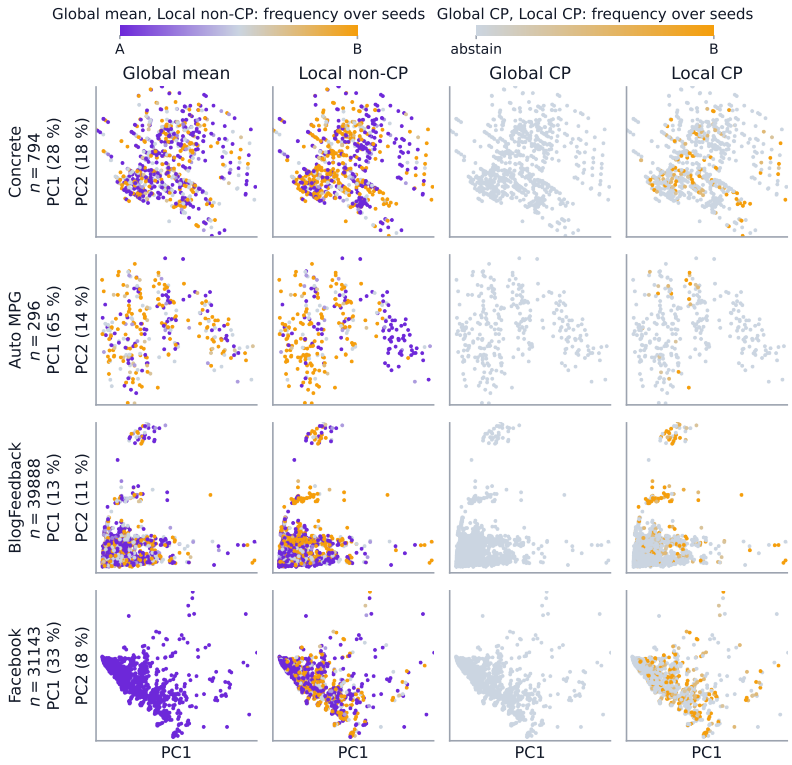

Figure 4: Regression winner regions in a 2D PCA projection. Colors show across-seed selection frequency: violet-to-amber for non-CP (A to B) and gray-to-amber for CP (abstention to B).

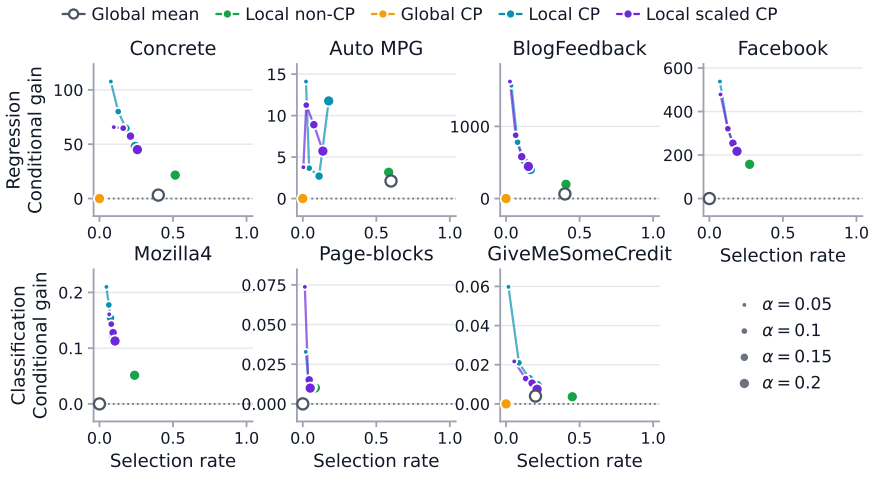

Figure 5: Conditional gain versus selection rate on four regression datasets (top) and three classification datasets (bottom). Larger markers correspond to larger α.

8 Limitations

The independent three-way split $(I_{\text{fit}}, I_{\text{est}}, I_{\text{cal}})$ is clean but sample-inefficient; existing data-recycling methods can be explored (Vovk, 2015; Barber et al., 2021). The guarantee is marginal for the realized future score; exact conditional validity requires further assumptions (Foygel Barber et al., 2021). Near tie boundaries, abstention may limit recovery of weak local advantages. Simultaneous error control over an entire covariate map needs further treatment. Broader benchmarks and larger model suites would help demonstrate practical utility.

AI USE STATEMENT

Large language models were used for writing and editing assistance only. They are not a component of the proposed method, the theoretical results, or the experimental pipeline, and no reported result depends on them. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is conducted entirely on synthetic data and open tabular benchmarks without human subjects or novel data collection. Although local model comparisons can inform model routing in sensitive settings such as credit assessment, real-world deployment requires rigorous evaluation of subgroup fairness, distribution shifts, and licensing constraints. The predictive guarantees established here govern comparison errors alone and should not be extrapolated beyond this context.

REPRODUCIBILITY STATEMENT

Anonymized code, including the dataset loader and processed datasets, the localizers, the experiment runners and the scripts that generate every figure and table, is provided in the supplementary material. The assumptions are stated with each result, and complete proofs are given in Appendix A. The synthetic designs are described in Appendix B.1, the evaluation metrics and seeds in Appendix B.2, and the real datasets and fitted predictors in Appendix B.4. The localizers and the training of the TabPFN head are described in Section 4.2 and Appendix C. The data processing steps are documented in the dataset README of the code. Full reruns additionally require the separately licensed TabPFN-3.5 checkpoint.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6679–6687, 2021.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. 2019.

Yu Bai, Song Mei, Huan Wang, Yingbo Zhou, and Caiming Xiong. Efficient and differentiable conformal prediction with general function classes. *arXiv preprint arXiv:2202.11091*, 2022.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Nicolo Colombo and Vladimir Vovk. Training conformal predictors. In *Conformal and Probabilistic Prediction and Applications*, pp. 55–64. PMLR, 2020.

Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in neural information processing systems*, 35:22380–22395, 2022.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data. *Advances in Neural Information Processing Systems*, 38, 2026.

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.

Raffaella Giacomini and Halbert White. Tests of conditional predictive ability. *Econometrica*, 74(6): 1545–1578, 2006.

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

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

David Holzmüller, Léo Grinsztajn, and Ingo Steinwart. Better by default: Strong pre-tuned mlps and boosted trees on tabular data. In *Neural Information Processing Systems*, 2024.

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.

Benjamin Jäger, Nick Erickson, Léo Grinsztajn, Felix Birkel, Klemens Flöge, Oscar Key, Kürşat Kaya, Jonas Kübler, Adèle Frankel, Tobias Schröder, et al. Tabpfn-3.5: Technical report. *arXiv preprint arXiv:2609.17895*, 2026.

Ying Jin and Emmanuel J Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

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

Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after data-dependent model selection. *Journal of the American Statistical Association*, (just-accepted):1–26, 2026.

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

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, Christian A Naesseth, and Eric Nalisnick. Max-rank: Efficient multiple testing for conformal prediction. *arXiv preprint arXiv:2311.10900*, 2023.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74 (1):9–28, 2015.

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

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Yuhao Wang and Tengyao Wang. Localized conformal model selection. *arXiv preprint arXiv:2602.19284*, 2026.

A Appendix: Proofs

A.1 Proof of Proposition 1

Remark 1 (Validity is agnostic to the score function)**.** *A basic property of split conformal prediction (Vovk et al., 2005; Lei et al., 2018) is that finite-sample marginal coverage relies on exchangeability of the calibration and test residuals. The theory imposes no regularity, continuity, or boundedness assumptions on the comparison score $S(X, Y)$ itself. Because the centering and scaling functions, $\widehat{\mu}$ and $\widehat{s}$, are fitted on a sample disjoint from the calibration and test sets, conditional exchangeability of the residuals is preserved. Consequently, the choice of score function and the choice of local centering or scaling affect statistical efficiency, such as selection rate and power, but not the finite-sample validity of the conformal bound.*

Proof of Proposition 1. Condition on the data in the fitting and estimation splits, denoted collectively by $\mathcal{D}_0$. After conditioning on $\mathcal{D}_0$, the fitted predictors, the localized center $\widehat{\mu}(\cdot)$, and the localized scale $\widehat{s}(\cdot)$ are fixed measurable functions. The remaining randomness comes only from the calibration observations $(X_i, Y_i)_{i \in I_{\text{cal}}}$ and the test point (X_{n+1}, Y_{n+1}).

Let

$$S_{n+1} \coloneqq S(X_{n+1}, Y_{n+1}) \quad \text{and} \quad R_{n+1} \coloneqq \frac{S_{n+1} - \widehat{\mu}(X_{n+1})}{\widehat{s}(X_{n+1})}.$$

Together with the calibration residuals R_i defined in the main text, the collection $\{R_i : i \in I_{\text{cal}}\} \cup \{R_{n+1}\}$ is exchangeable conditional on $\mathcal{D}_0$.

Let $n_{\text{cal}} = |I_{\text{cal}}|$ and $k_\alpha = \lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil$, as defined in Section 4. If $k_\alpha > n_{\text{cal}}$, then $\widehat{q}_{1-\alpha} = \infty$ and the desired coverage statement is immediate. Otherwise, $\widehat{q}_{1-\alpha} = R_{(k_\alpha)}$, the k_α-th order statistic of the calibration residuals. By exchangeability, using random tie-breaking only for the rank argument, the rank of R_{n+1} among the $n_{\text{cal}} + 1$ residuals is uniform on $\{1, \ldots, n_{\text{cal}} + 1\}$. Moreover, the event $\{R_{n+1} > R_{(k_\alpha)}\}$ implies that this rank is larger than k_α. Therefore,

$$\mathbb{P}\left(R_{n+1} > \widehat{q}_{1-\alpha} \mid \mathcal{D}_0\right) \leq \frac{n_{\text{cal}} + 1 - k_\alpha}{n_{\text{cal}} + 1} \leq \alpha,$$

or equivalently,

$$\mathbb{P}\left(R_{n+1} \leq \widehat{q}_{1-\alpha} \mid \mathcal{D}_0\right) \geq 1 - \alpha.$$

Because $\widehat{s}(x) > 0$ by construction, the event $\{R_{n+1} \leq \widehat{q}_{1-\alpha}\}$ is equivalent to

$$S_{n+1} \leq \widehat{\mu}(X_{n+1}) + \widehat{s}(X_{n+1})\widehat{q}_{1-\alpha} = \widehat{U}_\alpha(X_{n+1}),$$

where the last equality follows from the definition of the conformal upper bound. Marginalizing over $\mathcal{D}_0$ gives

$$\mathbb{P}\left(S_{n+1} \leq \widehat{U}_\alpha(X_{n+1})\right) \geq 1 - \alpha.$$

Finally, the false-winner event satisfies

$$\{\widehat{U}_\alpha(X_{n+1}) < 0,\ S_{n+1} \geq 0\} \subseteq \{S_{n+1} > \widehat{U}_\alpha(X_{n+1})\}.$$

Hence,

$$\mathbb{P}\left(\widehat{U}_\alpha(X_{n+1}) < 0,\ S_{n+1} \geq 0\right) \leq \mathbb{P}\left(S_{n+1} > \widehat{U}_\alpha(X_{n+1})\right) \leq \alpha.$$

□

A.2 Proof of Theorem 1

Proof of Theorem 1. Throughout the proof, we condition on the fitted predictors produced by the independent fitting split, so that the evaluation observations are i.i.d. and $\mu(x) = \mathbb{E}[S(X, Y) \mid X = x]$ is fixed. Define

$$N_n \coloneqq \frac{1}{n} \sum_{i=1}^{n} K_{r_n}(X_i, x_0) S_i, \qquad D_n \coloneqq \frac{1}{n} \sum_{i=1}^{n} K_{r_n}(X_i, x_0).$$

The ratio N_n/D_n is identical to the estimator in (2) whenever $D_n > 0$. On the event $D_n = 0$, we may assign the estimator any fixed value, say zero; the argument below shows that $\mathbb{P}(D_n > 0) \to 1$, so this finite-sample convention has no asymptotic effect.

Let $\kappa_0 := \int_{\mathbb{R}^d} K(u)\,du$. By the theorem's kernel condition, $0 < \kappa_0 < \infty$. Since K is compactly supported, there exists $R_K < \infty$ such that $K(u) = 0$ whenever $\|u\| > R_K$. Since p_X is continuous at x_0 and $p_X(x_0) > 0$, p_X is locally bounded near x_0 and remains positive on a sufficiently small neighborhood of x_0. Thus, for all large n, every point $x_0 + r_n u$ with $K(u) \neq 0$ lies in a neighborhood where the local continuity and moment conditions apply.

First consider the denominator. By the change of variables $u = (x - x_0)/r_n$,

$$\mathbb{E}[D_n] = \int_{\mathbb{R}^d} K(u) p_X(x_0 + r_n u)\,du.$$

Because p_X is continuous at x_0 and K is bounded with compact support, dominated convergence gives

$$\mathbb{E}[D_n] \longrightarrow p_X(x_0) \int_{\mathbb{R}^d} K(u)\,du = p_X(x_0)\kappa_0 =: d_0.$$

Since $p_X(x_0) > 0$ and $\kappa_0 > 0$, we have $d_0 > 0$.

Next consider the numerator. By iterated expectation,

$$\mathbb{E}[N_n] = \mathbb{E}[K_{r_n}(X, x_0)\mu(X)]\,.$$

Using the same change of variables,

$$\mathbb{E}[N_n] = \int_{\mathbb{R}^d} K(u)\mu(x_0 + r_n u) p_X(x_0 + r_n u)\,du.$$

The local second-moment condition and Jensen's inequality imply that μ is locally bounded near x_0:

$$|\mu(x)| \le \mathbb{E}[|S(X,Y)| \mid X = x] \le \mathbb{E}[S(X,Y)^2 \mid X = x]^{1/2}.$$

Together with the local boundedness of p_X, this gives an integrable dominating function proportional to $K(u)$. Since both μ and p_X are continuous at x_0, dominated convergence yields

$$\mathbb{E}[N_n] \longrightarrow \mu(x_0) p_X(x_0) \int_{\mathbb{R}^d} K(u)\,du = \mu(x_0) p_X(x_0)\kappa_0.$$

It remains to show that N_n and D_n concentrate around their expectations. For the denominator,

$$\mathrm{Var}(D_n) = \frac{1}{n}\mathrm{Var}(K_{r_n}(X, x_0)) \le \frac{1}{n}\mathbb{E}\big[K_{r_n}(X, x_0)^2\big]\,.$$

Again changing variables,

$$\mathbb{E}\big[K_{r_n}(X, x_0)^2\big] = r_n^{-d} \int_{\mathbb{R}^d} K(u)^2 p_X(x_0 + r_n u)\,du.$$

Since K is bounded and compactly supported, K^2 is integrable; since p_X is locally bounded at x_0, the integral is $O(1)$. Hence

$$\mathrm{Var}(D_n) = O\!\left(\frac{1}{n r_n^d}\right) \longrightarrow 0,$$

because $n r_n^d \to \infty$. Therefore

$$D_n - \mathbb{E}[D_n] \xrightarrow{p} 0.$$

For the numerator,

$$\mathrm{Var}(N_n) = \frac{1}{n}\mathrm{Var}(K_{r_n}(X, x_0)S) \le \frac{1}{n}\mathbb{E}\big[K_{r_n}(X, x_0)^2 S^2\big]\,.$$

By iterated expectation and the same change of variables,

$$\mathbb{E}\big[K_{r_n}(X, x_0)^2 S^2\big] = r_n^{-d} \int_{\mathbb{R}^d} K(u)^2 \mathbb{E}\big[S^2 \mid X = x_0 + r_n u\big]\, p_X(x_0 + r_n u)\,du.$$

For all large n, the compact support of K keeps $x_0 + r_n u$ inside the local neighborhood where $\mathbb{E}[S^2 \mid X = x]$ is bounded and where p_X is bounded. Therefore

$$\mathbb{E}\big[K_{r_n}(X, x_0)^2 S^2\big] = O(r_n^{-d}),$$

and hence

$$\mathrm{Var}(N_n) = O\left(\frac{1}{n r_n^d}\right) \longrightarrow 0.$$

Thus

$$N_n - \mathbb{E}[N_n] \xrightarrow{p} 0.$$

Combining the expectation limits and concentration bounds gives

$$D_n \xrightarrow{p} p_X(x_0)\kappa_0$$

and

$$N_n \xrightarrow{p} \mu(x_0) p_X(x_0)\kappa_0.$$

Because $p_X(x_0)\kappa_0 > 0$, we also have $\mathbb{P}(D_n > 0) \longrightarrow 1$. Consequently, by Slutsky's theorem,

$$\hat{\mu}_n(x_0) = \frac{N_n}{D_n} \xrightarrow{p} \mu(x_0).$$

It remains to translate consistency of the local mean estimator into consistency of the local winner. If $\mu(x_0) < 0$, then $\delta^\star(x_0) = A$, and

$$\mathbb{P}\Big(\hat{\delta}_n(x_0) \neq A\Big) = \mathbb{P}(\hat{\mu}_n(x_0) \geq 0)\,.$$

Moreover,

$$\{\hat{\mu}_n(x_0) \geq 0\} \subseteq \{|\hat{\mu}_n(x_0) - \mu(x_0)| \geq |\mu(x_0)|\}\,.$$

Since $\hat{\mu}_n(x_0) \xrightarrow{p} \mu(x_0)$, the probability of the event on the right converges to zero. Hence

$$\mathbb{P}\Big(\hat{\delta}_n(x_0) = A\Big) \longrightarrow 1.$$

The case $\mu(x_0) > 0$ is identical and gives

$$\mathbb{P}\Big(\hat{\delta}_n(x_0) = B\Big) \longrightarrow 1.$$

Therefore, whenever $\mu(x_0) \neq 0$,

$$\mathbb{P}\Big(\hat{\delta}_n(x_0) = \delta^\star(x_0)\Big) \longrightarrow 1.$$

The same proof applies to the neighborhood-average estimation-split instance described in Section 5 after replacing n by $|I_{\mathrm{est}}|$, with the corresponding radius sequence satisfying the same shrinking-neighborhood and diverging-effective-sample-size conditions. □

A.3 Proof of Proposition 2

Proof of Proposition 2. We first prove consistency of the aggregate selector for the sign of the global mean score. Let $S_i := S(X_i, Y_i)$. By definition, $\bar{S}_n = \frac{1}{n}\sum_{i=1}^n S_i$. Since $\mathbb{E}[|S(X,Y)|] < \infty$, the weak law of large numbers implies

$$\bar{S}_n \xrightarrow{p} \mathbb{E}[S(X,Y)] = \mu_{\mathrm{glob}}.$$

Suppose first that $\mu_{\mathrm{glob}} > 0$. Then the true aggregate winner is B, and

$$\mathbb{P}\left(\hat{\delta}_n^{\mathrm{agg}} = B\right) = \mathbb{P}(\bar{S}_n \geq 0).$$

Hence

$$\mathbb{P}(\bar{S}_n < 0) \leq \mathbb{P}\left(|\bar{S}_n - \mu_{\mathrm{glob}}| \geq \mu_{\mathrm{glob}}\right) \longrightarrow 0,$$

so

$$\mathbb{P}\left(\hat{\delta}_n^{\mathrm{agg}} = B\right) \longrightarrow 1.$$

The case $\mu_{\mathrm{glob}} < 0$ is identical and yields

$$\mathbb{P}\left(\hat{\delta}_n^{\mathrm{agg}} = A\right) \longrightarrow 1.$$

Therefore,

$$\mathbb{P}\left(\hat{\delta}_n^{\mathrm{agg}} = \delta^{\mathrm{agg},\star}\right) \longrightarrow 1.$$

We next prove the existence statement. Fix any $q \in (1/2, 1)$. Let $X \sim \mathrm{Unif}(0,1)$, and let $\mathcal{R}_A = [0, q]$. Define

$$\mu(x) = \begin{cases} -a, & x \in \mathcal{R}_A, \\ b, & x \notin \mathcal{R}_A, \end{cases}$$

where $a > 0$ and $b > qa/(1-q)$. Then

$$\mathbb{P}(\mu(X) < 0) = \mathbb{P}(X \in \mathcal{R}_A) = q.$$

Moreover,

$$\mathbb{E}[\mu(X)] = -q\,a + (1-q)b > 0.$$

Let η be any mean-zero noise, independent of X, with $\mathbb{E}[|\eta|] < \infty$ and $\mathrm{Var}(\eta) > 0$, and define $S \coloneqq \mu(X) + \eta$. Then $\mathbb{E}[|S|] < \infty$, $\mathbb{E}[S \mid X = x] = \mu(x)$, and the score is non-degenerate because $\mathrm{Var}(S \mid X = x) = \mathrm{Var}(\eta) > 0$. Therefore

$$\mathbb{P}(\mu(X) < 0) = q \qquad \text{and} \qquad \mathbb{E}[S] > 0.$$

Applying the aggregate selector to i.i.d. draws from this constructed distribution, the first part of the proof gives

$$\mathbb{P}\left(\hat{\delta}_n^{\mathrm{agg}} = B\right) \longrightarrow 1.$$

This proves that aggregate selection can asymptotically favor model B even though model A is locally superior on at least a q-fraction of the covariate distribution. □

A.4 Proof of Proposition 3

Proof of Proposition 3. Fix $x \in \mathcal{X}$. Conditional on $X = x$, write $Y = f^\star(x) + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X = x] = 0$. For a generic model $m \in \{A, B\}$,

$$\begin{aligned} L_m(X,Y) \mid X = x &= \big(\hat{f}_m(x) - f^\star(x) - \varepsilon\big)^2 \\ &= \big(\hat{f}_m(x) - f^\star(x)\big)^2 - 2\varepsilon\big(\hat{f}_m(x) - f^\star(x)\big) + \varepsilon^2. \end{aligned} \tag{4}$$

Taking expectation over both the training randomness and the fresh test noise, we isolate the cross-term. Because the fresh test noise ε is independent of the fitting sample $\mathcal{D}$ conditional on X, we have $\mathbb{E}[\varepsilon \mid X = x, \mathcal{D}] = \mathbb{E}[\varepsilon \mid X = x] = 0$. Therefore,

$$\begin{aligned} \mathbb{E}\big[\varepsilon \hat{f}_m(x) \mid X = x\big] &= \mathbb{E}\Big[\mathbb{E}\big[\varepsilon \hat{f}_m(x) \mid X = x, \mathcal{D}\big] \mid X = x\Big] \\ &= \mathbb{E}\Big[\hat{f}_m(x)\,\mathbb{E}\big[\varepsilon \mid X = x, \mathcal{D}\big] \mid X = x\Big] \\ &= 0. \end{aligned} \tag{5}$$

Consequently,

$$\mathbb{E}[L_m(X,Y) \mid X = x] = \mathbb{E}\big[(\hat{f}_m(x) - f^\star(x))^2\big] + \mathbb{E}[\varepsilon^2 \mid X = x]. \tag{6}$$

Applying the standard bias–variance identity $\mathbb{E}[(Z-a)^2] = \mathrm{Var}(Z) + (\mathbb{E}[Z] - a)^2$ with $Z = \hat{f}_m(x)$ and $a = f^\star(x)$ gives

$$\mathbb{E}[L_m(X,Y) \mid X = x] = \mathrm{Var}_m(x) + \mathrm{Bias}_m^2(x) + \mathbb{E}[\varepsilon^2 \mid X = x]. \tag{7}$$

Subtracting the identities for $m = A$ and $m = B$ cancels the irreducible noise term $\mathbb{E}[\varepsilon^2 \mid X = x]$, yielding

$$\mu_{\mathrm{ex}}(x) = \mathbb{E}_{\mathcal{D},Y|X=x}[S_{\mathcal{D}}(X,Y) \mid X = x] = \mathrm{Bias}_A^2(x) - \mathrm{Bias}_B^2(x) + \mathrm{Var}_A(x) - \mathrm{Var}_B(x).$$

□

A.5 JOINT CALIBRATION AND RANK INTERVALS FOR MULTIPLE MODELS

Pairwise comparisons give a natural extension to M fitted models, but separate marginal guarantees do not control the probability of any erroneous comparison (Timans et al., 2023). We consider one simple extension in which these comparisons are jointly calibrated to obtain a partial order. Index the models by $1, \dots, M$, with losses $L_1, \dots, L_M$. For an ordered pair (i, j) of distinct models, let $S_{ij} = L_i - L_j$, so that $S_{ij} < 0$ means that model i has the smaller loss. Its center $\widehat{\mu}_{ij}$ and positive scale $\widehat{s}_{ij}$ are fitted on I_{est}. The two directions of a pair share one fit, with $\widehat{\mu}_{ji} = -\widehat{\mu}_{ij}$ and $\widehat{s}_{ji} = \widehat{s}_{ij}$. For $t \in I_{\text{cal}} \cup \{n+1\}$, replace the calibration residual by

$$R_t = \max_{i \neq j} \frac{S_{ij}(X_t, Y_t) - \widehat{\mu}_{ij}(X_t)}{\widehat{s}_{ij}(X_t)}, \qquad \widehat{U}_{ij,\alpha}(x) = \widehat{\mu}_{ij}(x) + \widehat{s}_{ij}(x)\widehat{q}_{1-\alpha}, \tag{8}$$

using the same conformal order statistic $\widehat{q}_{1-\alpha}$. We declare that model i beats model j at x when $\widehat{U}_{ij,\alpha}(x) < 0$. Since $R_{n+1} \leq \widehat{q}_{1-\alpha}$ is equivalent to $S_{ij}(X_{n+1}, Y_{n+1}) \leq \widehat{U}_{ij,\alpha}(X_{n+1})$ for all ordered pairs, Proposition 1's rank argument gives

$$\mathbb{P}\Big\{\exists\, i \neq j :\ \widehat{U}_{ij,\alpha}(X_{n+1}) < 0,\ S_{ij}(X_{n+1}, Y_{n+1}) \geq 0\Big\} \leq \alpha. \tag{9}$$

The joint quantile handles multiplicity and pairwise dependence without requiring a Bonferroni adjustment. The resulting partial order produces rank intervals across all models, preserving pairwise comparisons even when a model's exact rank is not determined. A single ordered comparison (A, B) reduces to the two-model rule. Appendix B.6 gives a three-model illustration.

Condition on the fitting and estimation splits and on any fitting randomization. The predictors, centers, scales, and set of ordered comparisons are then fixed. The scalar residuals in (8) are exchangeable across the calibration observations and a future observation. Applying the proof of Proposition 1 to these maximum residuals gives $\mathbb{P}(R_{n+1} \leq \widehat{q}_{1-\alpha}) \geq 1 - \alpha$. Since every scale is positive,

$$R_{n+1} \leq \widehat{q}_{1-\alpha} \quad \Longleftrightarrow \quad S_{ij}(X_{n+1}, Y_{n+1}) \leq \widehat{U}_{ij,\alpha}(X_{n+1}) \quad \text{for all } i \neq j.$$

On this event, every returned edge has $S_{ij} < 0$. Any erroneous returned edge is therefore contained in $\{R_{n+1} > \widehat{q}_{1-\alpha}\}$, proving (9). Taking any subset of the returned edges, or adding their transitive consequences, preserves this guarantee.

Let $\mathcal{B}_j(x)$ be the set of models with a directed path to j and $\mathcal{W}_j(x)$ the set reachable from j. The rank bounds are

$$\underline{\rho}_j(x) = 1 + |\mathcal{B}_j(x)|, \qquad \overline{\rho}_j(x) = M - |\mathcal{W}_j(x)|.$$

When all edges are correct, each model in $\mathcal{B}_j$ has strictly smaller loss than j, and each model in $\mathcal{W}_j$ has strictly larger loss. Hence all realized ranks lie in their respective intervals simultaneously. The same conclusion holds under ties for any total ordering compatible with the loss order. The lower and upper ranks also equal the minimum and maximum position of j over total orderings consistent with the returned edges. If the returned edges contain a cycle, they are discarded and every model receives $[1, M]$; this cannot reduce coverage.

For $m \in \{1, \dots, M\}$, let $\mathcal{T}_m(X, Y)$ be the set of the m models with the smallest realized losses at (X, Y), under any total ordering compatible with the loss order, and define $\widehat{\mathcal{T}}_m^{\text{in}}(x) = \{j : \overline{\rho}_j(x) \leq m\}$ and $\widehat{\mathcal{T}}_m^{\text{out}}(x) = \{j : \underline{\rho}_j(x) \leq m\}$. On the same event of probability at least $1 - \alpha$, $\widehat{\mathcal{T}}_m^{\text{in}}(X) \subseteq \mathcal{T}_m(X, Y) \subseteq \widehat{\mathcal{T}}_m^{\text{out}}(X)$ for every m. Thus a model can be certified as top-m without fixing its precise rank. These sets follow directly from the calibrated pairwise comparisons.

B APPENDIX: ADDITIONAL EXPERIMENTAL DETAILS

This appendix provides full details for the experiments reported in Section 6. We include the data-generating processes, visual illustrations of the fitted models, evaluation metrics, detailed non-conformal winner-identification results, real-data settings, localizer comparisons, the three-model example, and the score-family ablation.

Table 2: Synthetic data-generating processes. All cases use $X \sim \mathrm{Unif}[0,1]$ and $Y = f^\star(X) + \varepsilon$, with $\varepsilon \mid X = x \sim N(0, \sigma^2(x))$. The oracle pattern describes the intended comparison; evaluation uses the post-fit oracle.

Case / fixture	$f^\star(x)$ and noise	Oracle pattern / role
1: `homog_a`	$2(x-0.5) + 0.75\,\mathrm{sign}(\sin(6\pi x))$; $\sigma = 0.1$	A wins nearly everywhere. Role: power baseline.
2: `nonlinear_tail`	$2x$ for $x \le 0.95$; $2x + 0.45\sin(2\pi(x-0.95)/0.05)$ otherwise; $\sigma = 0.30$	B wins globally; A can win in the right tail. Role: global–local mismatch.
3: `middle_island`	$2x$ outside $[0.42, 0.58]$; $2x + 1.20\sin(2\pi(x-0.42)/0.16)$ inside; $\sigma = 0.30$	A wins globally through a narrow island; B wins on much of x. Role: map recovery.
4: `validity_stress`	Same signal as Case 1; $\sigma = 0.40$	A wins globally; realized scores are noisy. Role: CP necessity.
5: `hetero_scale`	Same signal as Case 1; $\sigma(x) = 0.1$ for $x \le 0.5$ and 2.0 otherwise	A wins globally; score dispersion changes sharply. Role: local scaling.

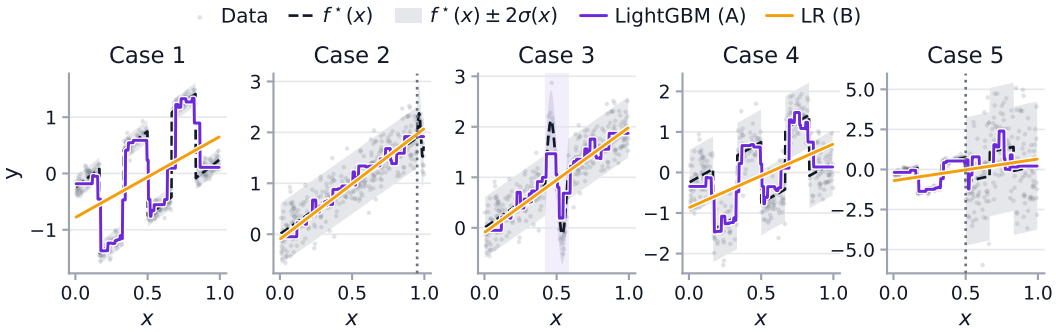

Figure 6: Synthetic DGPs. Black dashed: $f^\star(x)$; gray band: $f^\star(x) \pm 2\sigma(x)$; violet: LightGBM ($A$); amber: linear regression (B).

B.1 Synthetic data-generating processes

Table 2 reports the five synthetic data-generating processes. Model A is `LGBMRegressor` with 500 estimators, learning rate 0.05, and depth 6. Figure 6 visualizes the designs, together with representative fitted models. Case 1 is a baseline where A is better almost everywhere. Cases 2 and 3 create global–local mismatch: the former has a small nonlinear tail, and the latter a stronger nonlinear island. In Case 3, the island can make the global average favor A even though B is preferable over much of the input space. The actual oracle regions are computed from the fitted predictors and need not coincide exactly with the prescribed signal segments. Case 4 increases the noise level while keeping the Case 1 signal, creating a stress test for false declarations. Case 5 introduces strong heteroscedasticity, testing whether local scaling improves conformal efficiency.

B.2 Evaluation metrics for synthetic experiments

Table 3 defines the evaluation metrics, separated into non-conformal winner metrics that assess oracle map recovery and conformal winner metrics that assess reliability, selectivity, and realized gain. Let $\tau = 0.01\,\mathrm{sd}(S_{\mathrm{est}})$, $\mathcal{R}_A = \{x : \mu(x) < -\tau\}$, and $\mathcal{R}_B = \{x : \mu(x) > \tau\}$. The oracle label $w^\star$ is A on $\mathcal{R}_A$, B on $\mathcal{R}_B$, and a tie otherwise. Non-conformal rules return the sign-based winner, while a conformal rule returns A or abstains. Exact estimated ties are assigned to B in the experiment code.

Table 3: Evaluation metrics for synthetic experiments. Probabilities and expectations are estimated on the test split, conditional on the fitted models and localizer. For real data the declared model is B, and gains use $L_A - L_B$.

Metric	Definition	Interpretation
Non-conformal winner metrics		
Winner accuracy	$\mathbb{P}(\hat{w}(X) = w^\star(X) \mid \lvert\mu(X)\rvert > \tau)$	Agreement outside the oracle tie band
Region-IoU$_A$	$\mathbb{P}(\widehat{\mathcal{R}}_A \cap \mathcal{R}_A)/\mathbb{P}(\widehat{\mathcal{R}}_A \cup \mathcal{R}_A)$	Overlap with the oracle A region
Region-IoU$_B$	$\mathbb{P}(\widehat{\mathcal{R}}_B \cap \mathcal{R}_B)/\mathbb{P}(\widehat{\mathcal{R}}_B \cup \mathcal{R}_B)$	Overlap with the oracle B region
Conformal winner metrics		
Selection rate	$\mathbb{P}(\hat{w}_\alpha = A)$	Fraction of declared test points
False winner rate	$\mathbb{P}(\hat{w}_\alpha = A,\ S \geq 0)$	Marginal frequency of erroneous declarations
Power	$\mathbb{P}(\hat{w}_\alpha = A \mid X \in \mathcal{R}_A)$	Coverage of the oracle A region
Conditional gain	$\mathbb{E}[L_B - L_A \mid \hat{w}_\alpha = A]$	Average realized advantage in the selected region
Total gain	$\mathbb{E}[(L_B - L_A)\,\mathbb{I}_{\{\hat{w}_\alpha = A\}}]$	Realized gain per test observation

Winner accuracy excludes the tie band, whereas both IoUs use all test observations; points in the tie band are outside both oracle winner sets. This difference explains why high winner accuracy can coexist with more modest region overlap. Empty conditioning sets or empty IoU unions yield undefined values, which are omitted when averaging that metric across seeds. Conditional gain is also undefined when no point is selected, while total gain is then zero. The false winner rate uses all test observations as its denominator, including abstentions; it is not the error fraction conditional on declaring.

All reported tables use five seeds. Figure bands and table $\pm$ values are standard deviations across seeds, except for the Monte Carlo bars in Figure 3. In the dimension summary, cases are averaged within seed before computing the across-seed mean and standard deviation. For the score-family comparison in Appendix B.7, the oracle regions and gain units remain those of the raw squared-loss gap, including when the calibrated score is transformed.

B.3 Non-conformal winner-identification results

Table 4 reports local-winner identification results for the non-conformal winner maps. These results complement the main text by showing full accuracy and region-overlap values across all five synthetic cases.

B.4 Real datasets and fitted predictors

Table 5 summarizes the four regression and three binary classification datasets.

Every dataset is randomly partitioned into four parts: the first three contain $\lfloor n/4 \rfloor$ observations each and the remaining observations form the test set. Predictor features are standardized using the fitting split. A separate feature standardizer fitted on the estimation split is used for the localizer. Regression model B is ordinary least squares; classification model B is logistic regression with at most 2000 iterations and otherwise library defaults.

Table 4: Non-conformal local-winner identification, mean and standard deviation over five seeds. Bold entries mark the highest mean per metric within each case.

Case	Method	Accuracy	IoU_A	IoU_B
Case 1	Global mean	0.992 ± 0.004	0.992 ± 0.004	0.000 ± 0.000
	Local non-CP	0.995 ± 0.002	0.995 ± 0.002	0.400 ± 0.302
	Majority winner	$\mathbf{0.996 \pm 0.002}$	$\mathbf{0.996 \pm 0.002}$	0.420 ± 0.320
	Quantile winner	$\mathbf{0.996 \pm 0.003}$	$\mathbf{0.996 \pm 0.003}$	$\mathbf{0.569 \pm 0.227}$
Case 2	Global mean	0.941 ± 0.008	0.000 ± 0.000	0.691 ± 0.016
	Local non-CP	$\mathbf{0.956 \pm 0.019}$	$\mathbf{0.301 \pm 0.234}$	$\mathbf{0.761 \pm 0.035}$
	Majority winner	0.901 ± 0.066	0.182 ± 0.107	0.735 ± 0.022
	Quantile winner	0.941 ± 0.008	0.000 ± 0.000	0.691 ± 0.016
Case 3	Global mean	0.299 ± 0.020	0.151 ± 0.007	0.000 ± 0.000
	Local non-CP	$\mathbf{0.956 \pm 0.026}$	0.565 ± 0.100	0.448 ± 0.057
	Majority winner	0.922 ± 0.024	0.445 ± 0.033	$\mathbf{0.458 \pm 0.085}$
	Quantile winner	0.938 ± 0.025	$\mathbf{0.792 \pm 0.081}$	0.405 ± 0.052
Case 4	Global mean	0.992 ± 0.004	0.990 ± 0.007	0.000 ± 0.000
	Local non-CP	$\mathbf{0.993 \pm 0.004}$	$\mathbf{0.992 \pm 0.006}$	0.138 ± 0.248
	Majority winner	$\mathbf{0.993 \pm 0.004}$	$\mathbf{0.992 \pm 0.007}$	$\mathbf{0.204 \pm 0.258}$
	Quantile winner	0.414 ± 0.079	0.409 ± 0.081	0.013 ± 0.008
Case 5	Global mean	$\mathbf{0.856 \pm 0.023}$	$\mathbf{0.848 \pm 0.020}$	0.000 ± 0.000
	Local non-CP	0.804 ± 0.063	0.779 ± 0.070	0.299 ± 0.055
	Majority winner	0.796 ± 0.060	0.769 ± 0.069	$\mathbf{0.308 \pm 0.065}$
	Quantile winner	0.651 ± 0.021	0.593 ± 0.018	0.286 ± 0.047

Table 5: Real datasets: number of observations n and number of covariates d.

Task	Dataset	n	d
Regression	Concrete	1030	8
	Auto MPG	398	7
	BlogFeedback	52397	280
	Facebook	40948	53
Classification	Mozilla4	15545	4
	Page-blocks	5473	10
	GiveMeSomeCredit	30000	10

B.5 Detailed localizer comparison across dimensions

Table 6: Standardized conformal power by case, localizer, and dimension at $\alpha = 0.10$, with mean and standard deviation over five seeds. The last block averages cases within seed. Bold entries mark the highest case-specific or average mean in each column.

Case	Family	$d = 1$	$d = 10$	$d = 30$	$d = 100$
Case 1	k-NN	0.946 ± 0.026	0.967 ± 0.024	0.947 ± 0.044	0.958 ± 0.021
	Forest	0.826 ± 0.019	0.947 ± 0.016	0.928 ± 0.026	0.932 ± 0.024
	LightGBM	0.957 ± 0.020	0.783 ± 0.083	0.788 ± 0.093	0.792 ± 0.054
	MLP	0.975 ± 0.035	0.828 ± 0.046	0.938 ± 0.089	$\mathbf{0.991 \pm 0.018}$
	TabNet	0.933 ± 0.061	0.858 ± 0.055	0.871 ± 0.042	0.843 ± 0.081
	TabPFN	0.981 ± 0.021	0.974 ± 0.038	0.981 ± 0.016	0.977 ± 0.029
	TabPFN-gain	$\mathbf{0.991 \pm 0.005}$	$\mathbf{0.991 \pm 0.003}$	$\mathbf{0.989 \pm 0.003}$	$\mathbf{0.991 \pm 0.002}$

Continued on next page

Table 6 (continued)

Case	Family	$d = 1$	$d = 10$	$d = 30$	$d = 100$
Case 2	k-NN	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Forest	$\mathbf{0.174 \pm 0.139}$	$\mathbf{0.054 \pm 0.047}$	0.021 ± 0.018	0.000 ± 0.000
	LightGBM	0.019 ± 0.041	0.032 ± 0.027	0.020 ± 0.013	$\mathbf{0.064 \pm 0.013}$
	MLP	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	TabNet	0.000 ± 0.000	0.011 ± 0.024	$\mathbf{0.031 \pm 0.061}$	0.060 ± 0.123
	TabPFN	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	TabPFN-gain	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
Case 3	k-NN	0.510 ± 0.048	0.023 ± 0.022	0.014 ± 0.005	0.004 ± 0.007
	Forest	0.469 ± 0.085	0.557 ± 0.069	0.680 ± 0.062	$\mathbf{0.722 \pm 0.069}$
	LightGBM	0.563 ± 0.121	0.329 ± 0.139	0.275 ± 0.164	0.429 ± 0.057
	MLP	$\mathbf{0.710 \pm 0.073}$	0.002 ± 0.003	0.056 ± 0.038	0.014 ± 0.017
	TabNet	0.089 ± 0.200	0.237 ± 0.166	0.152 ± 0.066	0.082 ± 0.071
	TabPFN	0.532 ± 0.228	0.506 ± 0.095	0.664 ± 0.050	0.543 ± 0.188
	TabPFN-gain	0.636 ± 0.155	$\mathbf{0.616 \pm 0.038}$	$\mathbf{0.725 \pm 0.086}$	0.636 ± 0.091
Case 4	k-NN	0.133 ± 0.027	0.031 ± 0.021	0.030 ± 0.018	0.018 ± 0.010
	Forest	0.197 ± 0.017	0.125 ± 0.053	0.165 ± 0.036	$\mathbf{0.155 \pm 0.063}$
	LightGBM	0.158 ± 0.026	$\mathbf{0.176 \pm 0.037}$	0.153 ± 0.027	0.154 ± 0.022
	MLP	0.147 ± 0.066	0.052 ± 0.065	0.000 ± 0.000	0.000 ± 0.000
	TabNet	0.017 ± 0.030	0.095 ± 0.056	0.111 ± 0.048	0.065 ± 0.030
	TabPFN	0.116 ± 0.071	0.051 ± 0.075	0.119 ± 0.076	0.059 ± 0.071
	TabPFN-gain	$\mathbf{0.213 \pm 0.103}$	0.096 ± 0.113	$\mathbf{0.182 \pm 0.052}$	0.103 ± 0.072
Case 5	k-NN	0.500 ± 0.048	0.059 ± 0.008	0.007 ± 0.005	0.000 ± 0.000
	Forest	0.472 ± 0.027	$\mathbf{0.436 \pm 0.046}$	0.370 ± 0.049	0.267 ± 0.035
	LightGBM	$\mathbf{0.521 \pm 0.058}$	0.235 ± 0.029	0.190 ± 0.025	0.185 ± 0.012
	MLP	0.448 ± 0.103	0.194 ± 0.056	0.092 ± 0.055	0.001 ± 0.002
	TabNet	0.311 ± 0.084	0.087 ± 0.042	0.049 ± 0.024	0.058 ± 0.016
	TabPFN	0.392 ± 0.008	0.395 ± 0.023	0.404 ± 0.021	0.391 ± 0.024
	TabPFN-gain	0.414 ± 0.029	0.420 ± 0.029	$\mathbf{0.448 \pm 0.076}$	$\mathbf{0.411 \pm 0.040}$
Average (5 cases)	k-NN	0.418 ± 0.013	0.216 ± 0.010	0.200 ± 0.011	0.196 ± 0.003
	Forest	0.428 ± 0.032	0.424 ± 0.028	0.433 ± 0.024	0.415 ± 0.021
	LightGBM	0.443 ± 0.032	0.311 ± 0.036	0.285 ± 0.050	0.325 ± 0.017
	MLP	$\mathbf{0.456 \pm 0.009}$	0.215 ± 0.025	0.217 ± 0.030	0.201 ± 0.005
	TabNet	0.270 ± 0.058	0.258 ± 0.028	0.243 ± 0.032	0.221 ± 0.038
	TabPFN	0.404 ± 0.052	0.385 ± 0.021	0.434 ± 0.012	0.394 ± 0.044
	TabPFN-gain	0.451 ± 0.038	$\mathbf{0.425 \pm 0.022}$	$\mathbf{0.469 \pm 0.032}$	$\mathbf{0.428 \pm 0.027}$

Table 6 reports standardized conformal power separately for each case and localizer. TabPFN-gain improves on pretrained TabPFN in the five-case average at each dimension. In Case 3, TabPFN-gain has power 0.636 at both $d = 1$ and $d = 100$, while nearest-neighbor power falls from 0.510 to 0.004; in Case 5, TabPFN-gain retains power near 0.41 at both endpoints. Forests also perform well on Case 3 at high dimension. The detailed results reveal variation in efficiency across signal structures, with narrow-tail recovery remaining the most difficult case for standardized rules.

B.6 An exploratory comparison of three models

We apply Appendix A.5 to Case 1 with LightGBM, linear regression, and a one-hidden-layer MLP, again using five seeds. Figure 7 shows the rank intervals obtained from the jointly calibrated pairwise comparisons. Strongly separated models can receive exact ranks, while overlapping intervals retain unresolved comparisons. LightGBM receives rank 1 on most of the input range, while linear regression and the MLP share ranks 2 and 3 and swap order along x. This example illustrates how local comparison can retain partial ordering information without requiring a complete ranking at every input.

B.7 Score family

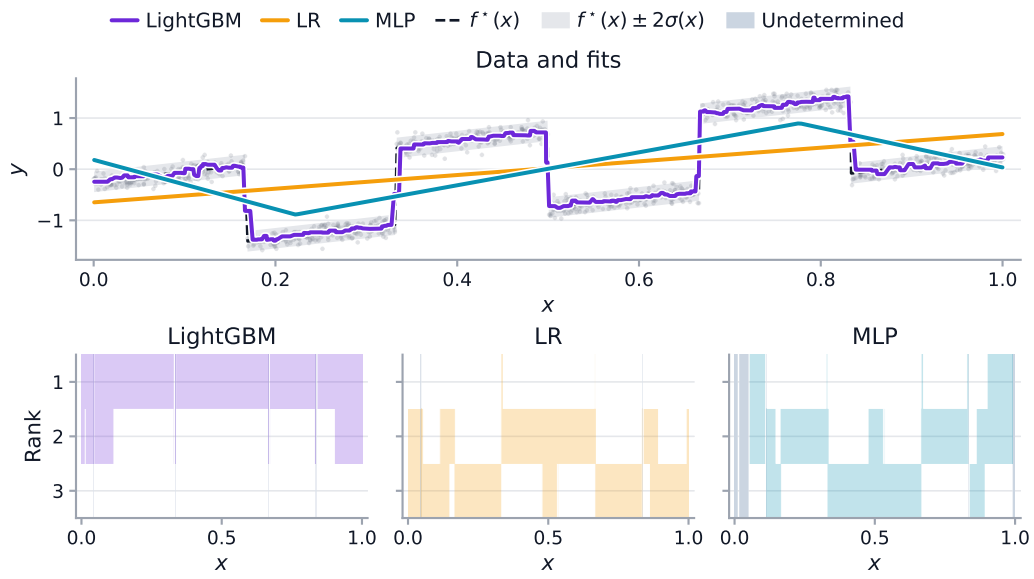

Figure 7: Three-model comparison on Case 1 at $\alpha = 0.10$. Top: test observations, signal, and fitted predictors. Bottom: realized-rank intervals from Local scaled CP for each model; narrower colored bands convey more information, while gray denotes the undetermined interval $[1, 3]$.

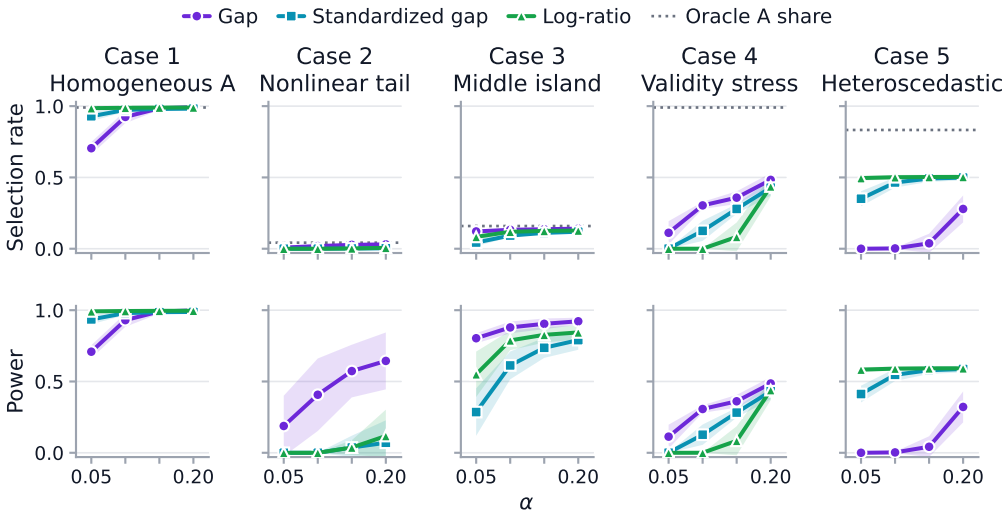

Figure 8: Score-family ablation using Local CP. Top: selection rate. Bottom: power within the raw squared-loss oracle A region. Gap, standardized gap, and log-ratio are each fitted and calibrated in their own score units. Curves and bands report mean and one standard deviation over five seeds.

Figure 8 compares three score choices: the squared-loss gap, a standardized gap, and a log-ratio of squared losses. This ablation uses the unstandardized local conformal rule in each score's own units. The log-ratio adds 10^{-8} to each squared loss before taking logarithms.

All three scores preserve the realized ordering of the two losses, so the same sign-based declaration and conformal validity argument apply. Their efficiency differs: the transformed scores improve power in the heteroscedastic Case 5, whereas the raw gap is more effective in the localized Cases 2 and 3. The oracle region in this figure is always defined by the raw squared-loss gap. This provides a common reference for comparing how score transformations affect local declarations.

C Appendix: Localizer Implementations

All localizers are fitted to the comparison score using only I_{est} and return a center and positive scale for calibration and test covariates. Their unstandardized two-model variants use unit scale. The two-model global baseline uses a constant center; its calibration-sample implementation gives the same bound as the raw score quantile because $\bar{S} + Q(S_i - \bar{S}) = Q(S_i)$. The multi-model global baseline instead fixes all centers and scales on I_{est} before joint calibration. The nearest-neighbor localizer averages the $k = \max(5, \lfloor\sqrt{|I_{\text{est}}|}\rfloor)$ closest estimation scores in standardized feature space with uniform weights, and uses their empirical standard deviation as the scale. The scale is floored at $\max(10^{-6}, 0.05\,\text{sd}(S_{\text{est}}))$.

The random-forest localizer fits a mean regressor and a variance regressor to squared out-of-bag residuals, using the default `RandomForestRegressor` settings. The LightGBM localizer fits analogous mean and variance models at library defaults, with squared residuals obtained by five-fold cross-fitting within I_{est}. These localizer trees differ from the 500-tree LightGBM predictor being compared with linear regression.

The MLP localizer has two hidden GELU layers. It is initialized by mean-squared-error training and then jointly fits the center and a log-variance head with a variance-weighted Gaussian loss, in which each observation's loss is weighted by its predicted variance raised to the power 0.5. Width is selected from $\{64, 128\}$ and weight decay from $\{10^{-4}, 10^{-3}, 10^{-2}\}$ by the Gaussian validation score on a fixed 20% stopping split within I_{est}. AdamW uses learning rate 10^{-3}, at most 1000 epochs per stage, and patience 50. This network is a location-scale baseline, distinct from RealMLP in the related literature and from the MLP predictors in the real-data comparisons.

TabNet fits one model for the mean score and a second for log squared residuals. Both use $n_d = n_a = 8$, three steps, $\gamma = 1.3$, sparsity penalty 10^{-3}, learning rate 0.02, batch size 128, virtual batch size 32, at most 300 epochs, and patience 30. The two networks use the same $80/20$ estimation/stopping split. Neural localizers standardize the score by its estimation-split mean and standard deviation and transform predictions back to score units. The network scale floor is 0.05 in standardized score units; TabNet and TabPFN scales are also capped at 100 in those units.

Pretrained TabPFN uses version 3.5 with four preprocessing ensemble members and a native distributional regression output. Its in-context reference set is the 80% training part of I_{est}; the remaining 20% is held out as a stopping split. Predictions use query batches of at most 8192 observations.

TabPFN-gain supplies the center and scale for the two local conformal methods, trained as in Section 4.2. Optimization uses AdamW with learning rate 10^{-5}, weight decay 10^{-3}, gradient clipping at 1, and at most 500 epochs. The dual step size is $\eta = 0.1$. Checkpoint selection uses hard total gain, averaged over the α grid and two fixed cross-fitted halves of the stopping split: one half calibrates and the other evaluates declarations; then their roles are exchanged. The selected checkpoint must satisfy both stopping-set risk caps and improve stopping gain; training-risk caps enter through the multiplier penalties. Stopping gain uses the α levels whose order statistics are finite in both stopping halves, so the smallest level can be omitted for small datasets. Patience is 50 epochs. The final head is loaded into the native TabPFN prediction path, with the same training context used for stopping evaluation. Both local conformal rules read the resulting center, and only the standardized rule uses its fitted scale.

The multiple-model experiment applies the same scalar training objective to each fitted pair, followed by joint maximum-residual calibration.