When Does Trimming Help Conformal Prediction? A Retained-Law Diagnostic under Calibration Contamination

Congye Wang[1,2]
[1]MARS, School of Mathematical Sciences, Lancaster University, UK
[2]School of Mathematics, Statistics & Physics, Newcastle University, UK
c.wang58@lancaster.ac.uk

Abstract

Trimming suspicious calibration points is a common response to contamination in conformal prediction. Its effect on clean-target coverage, however, is governed by the retained law induced by trimming, not by the contamination level alone. We analyse fixed-threshold trimming as conditioning rather than purification. It replaces the contaminated calibration law with a retained law, reducing clean-target coverage to a one-dimensional score-CDF transfer problem with an exact finite-sample identity. A componentwise bound on the transfer gap gives a population-level diagnostic. This separates a clean-side covariance cost from a retained-contamination cost, governed by the dirty-to-clean retention ratio. Trimming helps when the anomaly score separates retention probabilities while remaining score-neutral on the clean population. Otherwise, it cannot substantially reduce contamination through the retained mixture coefficient. We also give finite-sample certificate templates that provide numerical guarantees under independent audit.

1 Introduction

Trimming suspicious calibration points is a common response to contamination. A practitioner fits an anomaly score, such as an isolation, density, Stein, or geometric outlier score. Points with large anomaly values are then discarded, and ordinary split conformal prediction is applied to the retained sample. This procedure is intuitive and often believed to restore validity under calibration contamination. However, its effect on coverage depends on the retained law induced by trimming, as well as on properties of the anomaly score that cannot be inferred from the contamination level alone.

The difficulty is calibration–test exchangeability. Ordinary split conformal prediction is valid because, conditional on the fitted predictor and the nonconformity score, the calibration scores and the test score are exchangeable [Vovk et al., 2005, Shafer and Vovk, 2008, Lei et al., 2018, Angelopoulos and Bates, 2021]. Under calibration contamination, this exchangeability fails. The calibration sample is drawn from $P^\star = (1-\varepsilon)P + \varepsilon Q$, while the test point is drawn from the clean law P. Thus, calibration observations remain exchangeable among themselves, but not jointly exchangeable with the test point.

Trimming does not repair this mismatch. With a fixed threshold $t^\star$, it replaces $P^\star$ by the retained law

$$R = P^\star(\cdot \mid S \le t^\star), \tag{1}$$

which is generally neither $P^\star$ nor P. Conditional on the retained count, the retained observations are i.i.d. from R. Hence, conformal rank validity is exact only for a test point drawn from R, not from P. Clean-target validity is therefore a transfer problem from R back to P.

We turn this reframing into a population-level diagnostic for when trimming pays off. The transfer loss reduces to the supremum CDF gap between R and P under the nonconformity score. It also

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

admits a finite-sample lower bound that is sharp within the scalar-discrepancy information class. Substituting the retained-law mixture decomposition into this scalar loss gives a componentwise upper bound with two interpretable terms. The first is the clean-side cost of trimming. When S depends on the nonconformity score A under the clean law, conditioning on $\{S \le t^\star\}$ distorts the calibration distribution even without contamination. We show that this cost is exactly a one-sided covariance between the retention event and the score CDF under P. This formalises the design principle that a good anomaly score should be score-neutral on the clean population. The second term is the retained-contamination cost. Its coefficient is governed by the dirty-to-clean retention ratio. When the score separates clean and contaminated points, this coefficient can be made much smaller than ε. When the retention probabilities coincide, trimming cannot reduce contamination through the mixture coefficient. A favourable retained dirty score law is then only a separate, secondary mechanism. These two mechanisms organise both the theory and the experiments, through a score-visible regime and a no-separation regime.

This population diagnostic becomes a numerical guarantee only when its ingredients are bounded by independent information. We therefore complement it with finite-sample certificate templates. These templates provide coverage guarantees under independent audit, either through componentwise auditing of the mixture decomposition or through direct binomial inversion on an independent clean audit sample. We treat same-sample threshold selection separately, using a finite-grid empirical-CDF bound.

2 Background

Split conformal prediction. We work in the standard split conformal setting. Let $Z = (X, Y) \in \mathcal{Z}$, and let $A : \mathcal{Z} \to \mathbb{R}$ be a nonconformity score. The score is constructed using data independent of the final calibration set. The split-conformal prediction set at level α is

$$C_\alpha(x) = \{y : A(x, y) \le \hat{\tau}_\alpha\},$$

where $\hat{\tau}_\alpha$ is the inclusive empirical $(1-\alpha)$-quantile of the calibration scores. Once the model-fitting stage is fixed, validity reduces to a one-dimensional rank argument. The calibration scores and the test score are exchangeable. Hence, the resulting set covers Y with probability at least $1 - \alpha$ [Vovk et al., 2005, Shafer and Vovk, 2008, Lei et al., 2018, Angelopoulos and Bates, 2021]. Regression-oriented variants, such as conformalised quantile regression, modify the score construction. They nevertheless preserve the same calibration-quantile logic [Romano et al., 2019]. Other work seeks localised, weighted, or conditional guarantees, rather than the marginal clean-test guarantee considered here [Guan, 2023, Gibbs et al., 2025, Hore and Barber, 2025]. A growing literature studies failures of exchangeability caused by exogenous test-time distribution shift. Examples include covariate shift, label shift, time-series and online shift, non-exchangeability, and distributional robustness [Tibshirani et al., 2019, Podkopaev and Ramdas, 2021, Gibbs and Candes, 2021, Gibbs and Candès, 2024, Barber et al., 2023, Oliveira et al., 2024, Qiu et al., 2023, Yang et al., 2024, Zaffran et al., 2022, Bhatnagar et al., 2023, Aolaritei et al., 2025]. We study a different breakdown. It occurs on the calibration side and is induced endogenously by the practitioner's preprocessing.

Contamination model. The calibration observations are i.i.d. from $P^\star = (1-\varepsilon)P + \varepsilon Q$, where Q is an unknown contaminating law and $\varepsilon \in [0, 1)$ is the contamination level. The test point retains the clean law P. This is the classical Huber ε-contamination model [Huber, 1964], applied at the calibration stage. The contamination is non-adaptive. The law Q may be arbitrary, but it is not chosen with knowledge of the realised calibration sample or the trimming rule. The closest existing analysis is Clarkson et al. [2024]. They derive worst-case robust quantile corrections, which inflate the conformal threshold to absorb arbitrary Q. Their guarantee is uniform over Q. In contrast, our diagnostic conditions on the practitioner's chosen anomaly score and threshold. Their question is how much to inflate to be safe against any Q; ours is when trimming helps, given that the practitioner is already trimming. Adjacent work studies robust conformal validation under bounded corruption [Bashari et al., 2025, Cauchois et al., 2024a,b], local or cell-wise outlier models [Peng et al., 2025], and label-noise corrections for classification [Romano et al., 2020, Einbinder et al., 2024, Sesia et al., 2025, Penso and Goldberger, 2024, Bortolotti et al., 2025, Xi et al., 2025]. These target distinct corruption structures. Label-noise work, in particular, corrupts Y given X but preserves X. Our model imposes neither restriction. It isolates how trimming interacts with the conformal quantile itself.

Trimming. To trim, we introduce a second score $S : \mathcal{Z} \to \mathbb{R}$ for filtering. We also choose a threshold $t^\star \in \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. Calibration points are retained when $S(Z) \le t^\star$. Trimming has a long tradition in robust statistics, from Tukey's trimmed means to Huber's influence-function analysis [Huber, 1964]. In that setting, the target is an estimator, and trimming limits the influence of contaminating mass. Here, we study trimming as a transformation of the calibration law. The target is the conformal quantile. Trimming acts by conditioning the calibration law on a retention event, not by purifying it. The threshold is fixed after conditioning on auxiliary information. This may be a clean reference split, an independent tuning split, or external domain knowledge. The calibration sample therefore remains i.i.d. and independent of $t^\star$. Same-sample threshold selection is treated separately in Section 3.

Choice of anomaly score. We keep the score S general. It may be covariate-only, response-dependent, or externally supplied. The retained-law theorem treats it as an arbitrary fixed measurable function. In practice, the choice of S matters. Response-dependent filtering often couples retention with the nonconformity score under the clean law. This can inflate the clean-side distortion identified in Section 3. Our operational examples therefore favour covariate-geometric scores. For these scores, the clean distortion and the contamination-removal effect can be diagnosed separately. Section 3 formalises the central object of this paper, the retained law

$$R = P^\star(\cdot \mid S \le t^\star).$$

It then develops the diagnostic and certificate routes introduced above.

3 Methodology

We work conditionally on an auxiliary sigma-field $\mathcal{G}$. This captures the realised model-fitting and design stages. The fitted predictor, the nonconformity score A, the anomaly score S, and the trimming threshold $t^\star \in \overline{\mathbb{R}}$ are all $\mathcal{G}$-measurable.

The final calibration sample is i.i.d. from the contaminated law

$$P^\star = (1-\varepsilon)P + \varepsilon Q,$$

with $\varepsilon \in [0, 1)$. The clean test point is independent of the calibration sample and has law P.

All statements below are marginal in the clean test point and conditional on $\mathcal{G}$. We do not claim conditional, subgroup, or adversarial-contamination validity. Throughout, we distinguish population-level diagnostics from certificate templates. The latter yield numerical guarantees only when their ingredients are bounded by independent audit information.

3.1 Setup and the retained law

Given a fixed threshold $t^\star$, the trimmed procedure keeps calibration points with $S(Z_i) \le t^\star$. It then computes the inclusive $(1-\alpha)$-quantile of their nonconformity scores. Let N_{keep} denote the retained count. We set $\hat{\tau}_\alpha = +\infty$ if $N_{\mathrm{keep}} = 0$ or if

$$\lceil (N_{\mathrm{keep}} + 1)(1-\alpha) \rceil = N_{\mathrm{keep}} + 1.$$

Otherwise, $\hat{\tau}_\alpha$ is the $\lceil (N_{\mathrm{keep}} + 1)(1-\alpha) \rceil$-th retained order statistic. The prediction set is

$$C_\alpha(x) = \{y : A(x, y) \le \hat{\tau}_\alpha\},$$

For any probability law M on $\mathcal{Z}$, let M^A denote the distribution of the nonconformity score $A(Z)$ when $Z \sim M$. Define its cumulative distribution function by

$$F_M^A(a) := MA(Z) \le a.$$

We use the convention $F_M^A(+\infty) = 1$. For two probability laws M and N, define the one-sided score-CDF discrepancy

$$d^A_{M \to N,+} := \sup_{a \in \mathbb{R}} \big(F_M^A(a) - F_N^A(a)\big)_+,$$

where $(u)_+ = \max u, 0$. This discrepancy measures the largest positive difference between the score CDFs under M and N. The full trimming algorithm is given in Appendix B. The central object of our analysis is the *retained law*

$$R = P^\star(\,\cdot \mid S \le t^\star\,). \tag{2}$$

With clean and dirty retention probabilities $p_c = P(S \le t^\star)$ and $p_d = Q(S \le t^\star)$, and $P_{\mathrm{keep}}, Q_{\mathrm{keep}}$ the corresponding conditional laws when defined, Bayes' rule gives

$$R \;=\; (1-\tilde{\varepsilon}_\star)\,P_{\mathrm{keep}} \;+\; \tilde{\varepsilon}_\star\,Q_{\mathrm{keep}}, \qquad \tilde{\varepsilon}_\star \;=\; \frac{\varepsilon p_d}{(1-\varepsilon)p_c + \varepsilon p_d}. \tag{3}$$

We assume $\mu_{\mathrm{keep}} := (1-\varepsilon)p_c + \varepsilon p_d > 0$. If $p_d = 0$, Q_{keep} may be defined arbitrarily, and $\tilde{\varepsilon}_\star D_{Q,+}$ is interpreted as zero. If $p_c = 0$, the direct gap $d^A_{R\to P,+}$ remains meaningful, but componentwise quantities involving P_{keep} are not used. The *retained mixture coefficient* $\tilde{\varepsilon}_\star$ is not itself a measure of harmful contamination. Its effect on undercoverage is mediated by the product $\tilde{\varepsilon}_\star D_{Q,+}$.

Three score-law distances control the analysis. The first is the retained-to-clean discrepancy,

$$d^A_{R\to P,+} = \sup_t (F^A_R(t) - F^A_P(t))_+,$$

which governs clean-coverage transfer. The second is the clean trimming distortion,

$$\Delta^A_{\mathrm{trim},+} = \sup_t (F^A_{P_{\mathrm{keep}}}(t) - F^A_P(t))_+,$$

which measures how filtering perturbs the clean score CDF. The third is the retained dirty discrepancy,

$$D_{Q,+} = \sup_t (F^A_{Q_{\mathrm{keep}}}(t) - F^A_P(t))_+,$$

which measures how the retained contaminating component differs from the clean target.

Two-sided and downward variants are defined analogously. They are used for upper-coverage diagnostics in Appendix C.4.

3.2 Retained-law transfer and finite-sample sharpness

Conditional on $N_{\mathrm{keep}} = n$, the retained calibration scores are i.i.d. from R. Hence the trimmed cutoff is exchangeable with an independent test score from the retained law. This gives the standard rank guarantee under R. It then transfers to the clean target through the score-CDF gap.

Proposition 1 (Retained-law rank validity)**.** *Under the setup of Section 3.1 with $\mu_{\mathrm{keep}} > 0$, $\mathbb{E}\{F^A_R(\hat{\tau}_\alpha) \mid \mathcal{G}\} \ge 1-\alpha$, and for any independent target law P_0,*

$$\mathbb{P}_{Z_{\mathrm{new}}\sim P_0}\{A(Z_{\mathrm{new}}) \le \hat{\tau}_\alpha \mid \mathcal{G}\} \;\ge\; 1-\alpha-\sup_t\big(F^A_R(t) - F^A_{P_0}(t)\big)_+.$$

The marginal clean coverage has an exact finite-sample identity. Let

$$Q^A_M(u) := \inf\{a \in \mathbb{R} : F^A_M(a) \ge u\}$$

denote the lower generalised quantile of the score law M.

Theorem 1 (Exact finite-sample marginal coverage identity)**.** *Under the setup of Section 3.1 with $\mu_{\mathrm{keep}} > 0$, let $r_n = \lceil (n+1)(1-\alpha) \rceil$. Almost surely in $\mathcal{G}$,*

$$\mathbb{P}\{Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \mid \mathcal{G}\} \;=\; \sum_{n=0}^{m} \binom{m}{n} \mu^n_{\mathrm{keep}}(1-\mu_{\mathrm{keep}})^{m-n}\Psi_n(P,R), \tag{4}$$

where $\Psi_n(P,R) = 1$ if $r_n = n+1$, and $\Psi_n(P,R) = \mathbb{E}\big[F^A_P\{Q^A_R(B_{r_n,n+1-r_n})\}\big]$ otherwise, with $B_{r_n,n+1-r_n} \sim \mathrm{Beta}(r_n, n+1-r_n)$. The same identity holds for any independent target P_0 in place of P. The lower generalised inverse convention does not require continuity of F^A_R; see Appendix D, Lemma 1, for the atom-handling details.

The exact identity is unwieldy. Retaining only $d := d^A_{R\to P,+}$ gives the scalar lower bound

$$L_{\mathrm{fs}}(d) \;:=\; \sum_{n=0}^{m} \binom{m}{n} \mu^n_{\mathrm{keep}}(1-\mu_{\mathrm{keep}})^{m-n}\,\psi_n(d), \qquad \psi_n(d) = \mathbb{E}\big[(B_{r_n,n+1-r_n} - d)_+\big] \tag{5}$$

($\psi_n(d) = 1$ when $r_n = n+1$), obtained by replacing $F^A_P \circ Q^A_R$ in (4) by $u \mapsto (u-d)_+$.

Proposition 2 (Finite-sample scalar transfer)**.** *Under the setup of Theorem 1, almost surely in $\mathcal{G}$,*

$$\mathbb{P}\{Y_{\text{new}} \in C_\alpha(X_{\text{new}}) \mid \mathcal{G}\} \;\geq\; L_{\text{fs}}(d^A_{R\to P,+}) \;\geq\; \big[1-\alpha-d^A_{R\to P,+}\big]_+. \tag{6}$$

In closed form, $\psi_n(d) = \frac{r_n}{n+1}\{1-I_d(r_n+1, n+1-r_n)\} - d\{1-I_d(r_n, n+1-r_n)\}$ when $r_n \leq n$, with $I_x(a,b)$ the regularised incomplete beta function.

The scalar layer deliberately discards the joint structure of F^A_R and F^A_P. The next theorem makes the sharpness claim precise. Here, "best possible" means minimax optimal among uniform lower bounds that use only the scalar discrepancy constraint.

Theorem 2 (Finite-sample sharpness within the scalar information class)**.** *Fix $\alpha \in (0,1)$, $m \geq 0$, $\mu \in (0,1]$, and $d \in [0,1]$. For score laws R^A, P^A on $\mathbb{R}$, define $\mathcal{C}_{m,\mu}(R^A, P^A)$ as the clean-target coverage of the abstract retained-law conformal experiment: $N \sim \text{Binomial}(m,\mu)$; given $N=n$, $X_1,\dots,X_n \overset{\text{iid}}{\sim} R^A$ and independent $X_0 \sim P^A$; cutoff $+\infty$ if $r_n = n+1$, otherwise $X_{(r_n)}$. Let $\mathfrak{C}_d := \{(R^A,P^A) : \sup_t (F^A_R(t) - F^A_P(t))_+ \leq d\}$. Then*

$$\inf_{(R^A,P^A)\in\mathfrak{C}_d} \mathcal{C}_{m,\mu}(R^A,P^A) \;=\; L^{m,\mu}_{\text{fs}}(d) \;:=\; \sum_{n=0}^{m} \binom{m}{n} \mu^n (1-\mu)^{m-n}\, \psi_n(d), \tag{7}$$

attained by continuous R^A_d, P^A_d with $\sup_t (F^A_{R_d}(t) - F^A_{P_d}(t))_+ = d$. No uniform lower bound using only m, μ, α and $d^A_{R\to P,+} \leq d$ can exceed $L^{m,\mu}_{\text{fs}}(d)$. As $m\to\infty$ with μ fixed, $L^{m,\mu}_{\text{fs}}(d) \to [(1-\alpha)-d]_+$.

The extremal laws are constructed for the abstract scalar problem. Sharper bounds may be obtained from the mixture decomposition (3), smoothness of F^A_R, or audit information.

Substituting (3) into the definition of $d^A_{R\to P,+}$ and applying the triangle inequality gives

$$d^A_{R\to P,+} \leq (1-\tilde{\varepsilon}_\star)\Delta^A_{\text{trim},+} + \tilde{\varepsilon}_\star D_{Q,+}.$$

This is an upper bound, not an exact decomposition. Combined with Proposition 2, it yields

$$\mathbb{P}\{Y_{\text{new}} \in C_\alpha(X_{\text{new}}) \mid \mathcal{G}\} \geq \big[1-\alpha-(1-\tilde{\varepsilon}_\star)\Delta^A_{\text{trim},+} - \tilde{\varepsilon}_\star D_{Q+}\big]_+.$$

The first term is the clean-side conditioning cost. The second is the retained-contamination cost. The next two subsections analyse these two terms in turn.

3.3 Clean-side cost: a covariance identity

The clean trimming distortion captures how filtering perturbs the calibration distribution, even without contamination. Rewriting it as a covariance under P makes the mechanism clear.

Proposition 3 (Covariance identity for clean trimming distortion)**.** *Let $K = \mathbf{1}\{S(Z) \leq t^\star\}$ and assume $p_c > 0$. For every $t \in \mathbb{R}$,*

$$F^A_{P_{\text{keep}}}(t) - F^A_P(t) \;=\; \frac{\text{Cov}_P\big(\mathbf{1}\{A(Z)\leq t\},\, K\big)}{p_c}, \tag{8}$$

so $\Delta^A_{\text{trim},+} = p_c^{-1}\sup_t\{\text{Cov}_P(\mathbf{1}\{A(Z)\leq t\}, K)\}_+$, and $\Delta^A_{\text{trim}} = 0$ whenever $A(Z) \perp\!\!\!\perp K$ under P.

A useful anomaly score should be selective for contamination across P and Q, while remaining neutral with respect to score within P. To make this condition precise, consider the clean homoscedastic regression model $Y = f^\star(X) + \sigma\xi, \xi \perp X$, where $f^\star$ is the true regression function and $\widehat{f}$ is a fitted predictor. Let the nonconformity score be the absolute residual $A(X,Y) = |Y - \widehat{f}(X)|$. Suppose the retention indicator K depends only on X. If $\widehat{f} = f^\star$, then $A(X,Y) = A_0 := \sigma|\xi|$, which is independent of K under P. Consequently, $\Delta^A_{\text{trim}} = 0$. More generally, suppose that $\|f^\star - \widehat{f}\|_\infty \leq \eta$. The same argument gives

$$\Delta^A_{\text{trim}} \leq \omega_{A_0}(\eta) := \sup_t P\{A_0 \in [t-\eta, t+\eta]\}.$$

In particular, if A_0 has a density bounded above by M, then $\Delta^A_{\mathrm{trim}} \le 2M\eta$.

For regularised least-squares backbones, standard learning-theory controls can provide such prediction-error inputs under additional assumptions [Caponnetto and De Vito, 2007]. Our diagnostic only needs the realised bound. Response-dependent anomaly scores can make $\Delta^A_{\mathrm{trim},+}$ arbitrarily large, even without contamination. This motivates our use of covariate-geometric scores in the operational examples.

3.4 Retained mixture coefficient and retention-probability separation

The retained-contamination cost $\tilde{\varepsilon}_\star D_{Q,+}$ is governed by $\tilde{\varepsilon}_\star$, which depends only on the retention probabilities. If the retained dirty score law is harmless, so that $D_{Q,+} = 0$, the dirty contribution vanishes even when $\tilde{\varepsilon}_\star$ is large.

Proposition 4 (Coefficient-level separation dichotomy)**.** *Under the setup of Section 3.1 with $p_c > 0$:*

- *If $p_d \le \lambda p_c$, then $\tilde{\varepsilon}_\star \le \varepsilon\lambda/(1-\varepsilon+\varepsilon\lambda)$; in particular, $p_d/p_c \to 0$ forces $\tilde{\varepsilon}_\star \to 0$.*
- *If $p_d = p_c$, then $\tilde{\varepsilon}_\star = \varepsilon$.*

When the score induces *retention-probability separation*, dirty points are retained much less often than clean ones. Then $\tilde{\varepsilon}_\star$ can be far below ε, and trimming can help. Without separation, $\tilde{\varepsilon}_\star = \varepsilon$ for any threshold. Trimming then cannot reduce contamination through the mixture coefficient. Improvement can only come from a favourable Q_{keep}. Yet trimming still incurs the clean covariance cost in Proposition 3, so it can strictly worsen the diagnostic lower bound.

We call this the *coefficient-level separation dichotomy*. Section 4 organises the experiments around it. The proposition rules out improvement through the retained-contamination *coefficient*, not through all possible differences between Q_{keep} and Q. A stronger statement appears in Theorem 5.

3.5 Selection-aware bounds and certificate templates

Practical choice of the anomaly score and threshold. The anomaly score should be chosen according to the expected contamination mechanism. Covariate-based scores can be useful when contamination changes the distribution of X. They have no direct access to label-only corruption when $Q_X = P_X$.

For a candidate score S and threshold t, two quantities are especially important. The first is the clean trimming distortion $\Delta^A_{\mathrm{trim},+}(t)$. The second is the retained contamination coefficient

$$\widetilde{\varepsilon}(t) = \frac{\varepsilon p_d(t)}{(1-\varepsilon)p_c(t) + \varepsilon p_d(t)}.$$

A favourable rule should keep the clean score distribution nearly unchanged while making the ratio $p_d(t)/p_c(t)$ small.

When independent tuning information is available, a practical strategy is to consider a finite library of candidate scores and thresholds. Candidates can be screened using independently justified upper bounds on the population coverage loss. Among candidates that pass the coverage screen, one may choose the rule with the smallest estimated prediction-set size.

If no candidate passes the screen, the procedure returns to a calibration-independent fallback, such as no trimming. This avoids an unsupported trimming decision. However, the fallback does not automatically guarantee nominal clean-target coverage under contamination.

Theorem 3 justifies independent tuning without an additional selection penalty. Proposition 10 provides a separate bound for same-sample selection from a fixed finite threshold grid. The finite-library extension is given in Appendix F.5.

Same-sample threshold selection. Same-sample selection breaks the retained-law rank identity. What remains is a selection-uniform empirical-CDF transfer bound. For a finite grid $\mathcal{T} = \{t_1, \dots, t_K\}$, define N_t, R_t, $\hat{\tau}_t$, and $d_t = d^A_{R_t \to P,+}$ analogously. For $N_t \ge 1$, let

$$\eta_t = \sqrt{\log(2K/\beta)/(2N_t)}.$$

Then, with conditional probability at least $1-\beta$, the following holds simultaneously for all $t \in \mathcal{T}$ with $N_t \geq 1$ and $r_t \leq N_t$:

$$\mathbb{P}_P\{A(Z_{\text{new}}) \leq \hat{\tau}_t \mid Z_1, \ldots, Z_m, \mathcal{G}\} \geq \frac{r_t}{N_t} - d_t - \eta_t.$$

Proposition 10 in Appendix F gives the precise statement. The bound holds uniformly over the grid, allowing any data-dependent $\widehat{t} \in \mathcal{T}$ with a $\sqrt{\log K/N_t}$ penalty. It is data-driven only through the empirical-CDF term. It becomes a numerical certificate only after the population losses d_t are independently bounded. This view is related to conformal selection and adaptive novelty detection [Jin and Candès, 2023, Marandon et al., 2024], but here the selected object is the calibration subset used to form the conformal quantile. The same argument extends to a finite library of score-threshold pairs fixed before the final calibration sample is observed. The resulting bound incurs a $\sqrt{\log H/N_h}$ term. This extension permits data-dependent selection from a prespecified library, but does not cover unrestricted fitting of a new anomaly score on the final calibration sample. See Corollary 3 in Appendix F.5.

Procedure, diagnostic, and certificate. We distinguish three objects. Algorithm 1 constructs a prediction set using the nonconformity score A, the anomaly score S, a fixed threshold $t^\star$, and the contaminated calibration sample. It does not require knowledge of P, Q, the contamination level, or the latent contamination labels. Theorem 1 provides an exact population characterisation of clean target coverage. The component-wise diagnostic further explains the roles of clean trimming distortion and retained contamination. These population quantities describe the behaviour of a proposed trimming rule. They are not inputs to Algorithm 1. Numerical coverage certificates require additional information. Proposition 11 converts independently justified component bounds into a clean coverage guarantee. Proposition 12 instead uses an independent, cleanly labelled audit sample to certify the final prediction set.

Under unrestricted Huber contamination, clean target coverage cannot generally be identified from contaminated observations alone. Thus, the additional information required for numerical certification reflects an information limit of the model, rather than a missing step in Algorithm 1.

We consider two certificate routes. The first audits the mixture components. Given $L_c \leq p_c$, $p_d \leq U_d$, $\Delta^A_{\text{trim},+} \leq B_{\Delta,+}$, $D_{Q,+} \leq B_{Q,+}$, and $\varepsilon \leq \varepsilon_{\max}$,

$$\mathbb{P}\{Y_{\text{new}} \in C_\alpha(X_{\text{new}}) \mid \mathcal{G}\} \geq \left[1 - \alpha - B_{\Delta,+} - \bar{\varepsilon} \max\{B_{Q,+} - B_{\Delta,+}, 0\}\right]_+,$$

where

$$\bar{\varepsilon} = \frac{\varepsilon_{\max} U_d}{(1-\varepsilon_{\max})L_c + \varepsilon_{\max} U_d}.$$

The worst-case bound $B_{Q,+} = 1$ is always valid, but conservative.

The second route audits the final selected set directly, using an independent clean audit sample. Exact one-sided binomial inversion certifies the realised coverage. A one-sided Kolmogorov audit certifies any $\mathcal{A}$-measurable cutoff, and applies to any selected-subset procedure. Formal statements are given in Propositions 11 and 12 in Appendix F.

The diagnostics should be read as *simulation-evaluable diagnostics* when P and Q are known. Otherwise, they should be read as *certificate templates*. Numerical certification requires either independent component bounds or an independent audit sample. We maintain this distinction throughout Section 4.

4 Experimental Results

4.1 Setup and reporting conventions

We evaluate the retained law analysis through controlled regression experiments and semi-synthetic classification studies. The regression experiments isolate the two mechanisms in Section 3, clean trimming distortion and retained contamination. The classification experiments use real US Census covariates and labels with controlled cross state contamination. We also study corruption that affects labels only, the allocation of a limited clean data budget, and the agreement between theoretical

diagnostics and empirical performance. These experiments include both favourable and unfavourable regimes for trimming. The clean design is

$$X \sim N(0,1), \qquad Y = X + 0.6(1 + 0.6|X|)\xi, \qquad \xi \sim N(0,1),$$

and the contaminating distribution has

$$X_{\text{dirty}} \sim N(6,1), \qquad Y_{\text{dirty}} = X_{\text{dirty}} + 0.05\xi.$$

The contamination level is $\varepsilon = 0.2$. The contaminating responses are close to the regression surface. However, the contaminating covariates are clearly separated from the clean support. Thus, a covariate-geometric anomaly score should be able to detect them. Importantly, the clean design is heteroscedastic. Therefore, the homoscedastic-residual orthogonality in Proposition 3 does *not* apply, and nonzero clean trimming distortion is expected.

The nominal miscoverage level is $\alpha = 0.1$ throughout, so the clean coverage target is 0.9. Unless stated otherwise, the synthetic regression experiments use $m = 320$ calibration observations and 100 Monte Carlo repetitions. The ACS classification and clean budget experiments use 50 repetitions, as specified in their experimental settings. We report empirical clean coverage and prediction set width or size separately from population diagnostics and independently justified certificate bounds. The population quantities are evaluated using the known simulation mechanisms. They explain the observed behaviour but are not themselves data only certificates. The prediction rule and the Stein score-norm anomaly score are both fitted on independent clean auxiliary data. Thresholds are either fixed population thresholds or chosen on an independent reference split. No main-table row selects the threshold from the calibration sample used for the conformal quantile.

We report three types of quantities. *Empirical* clean coverage and width describe finite-sample performance, with Monte Carlo intervals across repetitions. *Population diagnostics*—p_c, p_d, $\tilde{\varepsilon}_\star$, $\Delta^A_{\text{trim},+}$, and $D_{Q,+}$—are evaluated from the known simulation laws. They are mechanistic explanations, not certificates. *Conservative lower bounds* ("Cons. LB") are componentwise audit-style certificates. They use independent diagnostic samples and concentration corrections, so they are operationally safe but deliberately loose.

When $D_{Q,+}$ is available from the simulation, we also report the population diagnostic

$$L_{\text{mix},+} = 1 - \alpha - (1 - \tilde{\varepsilon}_\star)\Delta^A_{\text{trim},+} - \tilde{\varepsilon}_\star D_{Q,+}.$$

This is the lower bound induced by the mixture upper bound on the scalar transfer loss. When the retained contaminating component is too rare to estimate reliably, we use the worst-case fallback $D_{Q,+} \le 1$.

4.2 Score-visible separation and the main trade-off

Appendix H.12 reports additional checks. As the Stein threshold rises from $q = 0.950$ to $q = 0.990$, the clean side covariance envelope falls from 0.0130 to 0.0036. It also reports the finite sample scalar bound curve, and Theorem 2 establishes its sharpness.

Combining the two mechanisms, clean-side covariance and retained contamination, Table 1 reports the main score-visible trade-off. Ordinary split conformal under-covers the clean target, with mean coverage 0.8709. All three Stein thresholds make dirty retention negligible, so $\tilde{\varepsilon}_\star$ is essentially zero in the trimmed rows. Coverage improves as q increases. At $q = 0.990$, mean coverage is 0.8950 (unrounded 0.89499), still below the nominal target 0.9, but close to the clean-calibration benchmark 0.8984. By Table 14, the improvement comes from reduced clean covariance distortion, not stronger contamination removal. The retained-contamination axis was already saturated at $q = 0.950$.

The $q = 0.990$ row also shows how the componentwise diagnostic separates the two mechanisms. The retained dirty discrepancy is $D_{Q,+} = 0.31$, but the retained mixture coefficient is only $\tilde{\varepsilon}_\star = 6.4 \times 10^{-5}$. Thus the dirty contribution is $\tilde{\varepsilon}_\star D_{Q,+} \approx 2.0 \times 10^{-5}$, which is dominated by the clean contribution $(1 - \tilde{\varepsilon}_\star)\Delta^A_{\text{trim},+} \approx 0.0036$. The resulting population diagnostic is $L_{\text{mix},+} = 0.8964$, close to the empirical mean 0.8950 within Monte Carlo error. The retained-law mixture upper bound therefore explains where the residual under-coverage comes from.

Method	Clean coverage	Width	p_c	p_d	$\tilde{\varepsilon}_\star$	$\Delta^A_{\mathrm{trim},+}$	$L_{\mathrm{mix},+}$	Cons. LB
Ordinary split	$0.8709\ [0.8667, 0.8750]$	2.7308	1.0000	1.0000	0.2000	0.0000	≥ 0.7000	0.5499
Stein $q = 0.950$	$0.8857\ [0.8819, 0.8896]$	2.8584	0.9494	2.8×10^{-5}	7.0×10^{-6}	0.0137	≥ 0.8863	0.7140
Stein $q = 0.975$	$0.8914\ [0.8877, 0.8951]$	2.9107	0.9752	6.1×10^{-5}	1.6×10^{-5}	0.0075	≥ 0.8925	0.7225
Stein $q = 0.990$	$0.8950\ [0.8913, 0.8986]$	2.9448	0.9897	2.6×10^{-4}	6.4×10^{-5}	0.0036	0.8964	0.7276
Clean oracle	$0.8984\ [0.8948, 0.9021]$	2.9791	1.0000	0.0000	0.0000	0.0000	0.9000	0.7322

Table 1: Score-visible regime with $\varepsilon = 0.2$, $m = 320$, and 100 repetitions. $L_{\mathrm{mix},+}$ is the population diagnostic lower bound. Entries marked $\geq$ use the worst-case fallback $D_{Q,+} \leq 1$. "Cons. LB" is a separate componentwise audit-style certificate. Trimming improves coverage from 0.8709 to 0.8950. The unrounded mean for $q = 0.990$ is 0.89499, so this row should be read as near-nominal and within Monte Carlo error of the clean oracle 0.8984, not as a formal 0.9 certificate. Width intervals are omitted for compactness; full intervals are given in Appendix H.

Regime	Method	Clean coverage	p_c	p_d	$\tilde{\varepsilon}_\star$	Δ_+	$D_{Q,+}$	$L_{\mathrm{mix},+}$	Cons. LB$_+$
Score-visible ($\varepsilon = 0.2$)	Stein $q = 0.990$	0.8950	0.9897	2.6×10^{-4}	6.4×10^{-5}	0.0036	0.31	0.8964	0.7277
Perfect dirty rejection	Stein clean-ref $q = 0.950$	0.9015	0.9444	0.0000	0.0000	0.0040	N/A	0.8960	0.7136
No separation, $m = 800$	Stein clean-ref $q = 0.950$	0.8754	0.9420	0.9429	0.2002	0.0037	0.94	0.7081	0.6028
Label-only, $\varepsilon = 0.20$	Ordinary split	0.9998	1.0000	1.0000	0.2000	0.0000	8.7×10^{-5}	0.9000	0.7631
Large $m = 10{,}000$	Stein $q = 0.950$	0.8888	0.9495	0.0057	0.0015	0.0134	0.86	0.8853	0.8464

Table 2: Diagnostic regimes. The score-visible row is the main result of Section 4.2. The no-separation row tests Proposition 4. The label-only and large-m rows show distinct failure modes.

4.3 The dichotomy in failure modes

Proposition 4 predicts that, without retention-probability separation, trimming cannot reduce $\tilde{\varepsilon}_\star$ through the mixture coefficient. Table 2 examines four regimes that stress this dichotomy from different sides.

Perfect dirty rejection. Placing the contamination cluster far from the clean support drives p_d to zero. The remaining loss is pure clean trimming distortion, with $\Delta^A_{\mathrm{trim},+} = 0.0040$. Coverage reaches 0.9015, slightly above nominal. This is because the nonzero clean retention probability, $p_c = 0.94$, does not perturb the conformal quantile in the under-coverage direction.

No separation. When clean and dirty points are retained with similar probabilities ($p_c = 0.942$, $p_d = 0.943$), Proposition 4 predicts $\tilde{\varepsilon}_\star \approx \varepsilon$. Empirically, $\tilde{\varepsilon}_\star = 0.200$, and the dirty contribution dominates, giving coverage 0.8754. Trimming does not help. Although the componentwise diagnostic is conservative because $D_{Q,+}$ is large, $\tilde{\varepsilon}_\star \approx \varepsilon$ confirms the no-separation mechanism. A separation sweep with $\varepsilon = 0.3$ and $m = 800$ shows the transition. As separation increases, p_d drops from 1.00 to 0.00, $\tilde{\varepsilon}_\star$ drops from 0.300 to 0.000, and coverage rises from 0.859 to 0.902. The improvement comes from retention separation, not trimming *per se*.

Label-only contamination. When only the response is corrupted ($Q_X = P_X$), a covariate-geometric score has no leverage. Ordinary split conformal over-covers (0.9998) by producing inflated intervals. Trimming with a covariate score cannot improve this, because the corrupted points are not detectable from X alone. This is a genuine limitation of covariate trimming, not a failure of the retained-law analysis.

Large calibration size. In the separate large calibration experiment, the empirical clean coverage is 0.8888 at $m = 10{,}000$, while the population diagnostic lower bound is $L_{\mathrm{mix},+} = 0.8853$. The empirical coverage lies above this lower bound, as expected. As the calibration size increases from 320 to 10,000, the conservative audit style lower bound rises from 0.7132 to 0.8464. In contrast, the population mixture diagnostic remains approximately unchanged within this experiment. These results show that increasing the calibration size reduces finite sample uncertainty in the certificate but does not by itself, remove the retained to clean score law discrepancy. The population diagnostic describes a potential coverage loss that persists beyond finite sample calibration effects. Further details are given in Appendix H.4.

ε	Ordinary APS coverage / size	Trimming coverage range	Trimming size range	$\tilde{\varepsilon}$ range
0.1	0.9004 / 1.4432	0.9003–0.9006	1.4430–1.4438	0.0998–0.1003
0.2	0.9006 / 1.4438	0.9003–0.9004	1.4429–1.4432	0.1997–0.2006
0.3	0.8992 / 1.4393	0.8989–0.8992	1.4384–1.4393	0.2996–0.3008

Table 3: Semi-synthetic ACS classification with California as the clean target and an equal mixture of Texas, New York, and Florida as contamination. Results are based on 50 Monte Carlo repetitions. Trimming ranges cover Isolation Forest, Mahalanobis, and GMM-Mahalanobis. Similar clean and dirty retention probabilities lead to little change in retained contamination or empirical coverage.

Threshold selection and audit. When the threshold is selected from the same calibration sample over a finite grid, Proposition 10 applies with a $\sqrt{\log K/N}$ penalty. When independent clean labelled audit data are available, Proposition 12 certifies the final selected set directly by binomial inversion, regardless of how the set was constructed. Across the tested policies, fixed and clean-reference $q = 0.950$ rules give coverage near 0.892. A more conservative KSD-geometric policy gives coverage 0.898 with larger width. Thus, the policy choice changes the coverage-efficiency frontier, but not the mechanism. Non-Gaussian geometries are reported in Appendix H. The qualitative conclusion is that the anomaly score should match the clean support geometry while remaining score-neutral within P, rather than being a universal choice.

4.4 Semi-synthetic ACS classification

We also evaluate the diagnostic on the 2018 American Community Survey (ACS) ACSEmployment classification task using Folktables [Ding et al., 2021]. California (CA) defines the clean target distribution P. Texas (TX), New York (NY), Florida (FL), and their equal mixture define alternative contaminating distributions Q. We consider contamination levels $\varepsilon \in \{0.1, 0.2, 0.3\}$. All methods use the same logistic regression predictor and randomised adaptive prediction sets (APS) score. We compare ordinary APS, direct clean calibration, source oracle trimming, and three implementable covariate anomaly scores: Isolation Forest, Mahalanobis, and GMM Mahalanobis. Source labels are used only to construct the contamination and evaluate the oracle. They are not supplied to the implementable trimming rules. Table 3 reports the results for the heterogeneous mixture. Across the three implementable anomaly scores, clean retention probabilities range from 0.9490 to 0.9502, while dirty retention probabilities range from 0.9477 to 0.9527. The retained contamination coefficient therefore remains close to the original contamination level. All practical trimming rules achieve coverage close to ordinary APS. Across all 12 combinations of contamination source and level, the largest absolute paired mean coverage change is only 0.0005. This is an informative negative result. Ordinary APS already achieves near nominal clean coverage in this setting. Moreover, common covariate anomaly scores do not meaningfully separate the clean and contaminating domains. These findings support the coefficient level prediction of Proposition 4. Trimming cannot substantially reduce the retained contamination coefficient when $p_c \approx p_d$. The experiment also illustrates an important limitation, successful trimming in a controlled, score visible setting does not imply similar benefits under realistic domain overlap.

5 Discussion

The retained-law viewpoint treats trimming as a transformation of the calibration law. Fixed-threshold trimming does not recover P, but induces a retained law R, under which conformal validity is exact. The cost is described by a one-dimensional score-CDF transfer. Trimming helps when the anomaly score separates retention under P and Q, while remaining nearly score-neutral within P. Otherwise, contamination is not reduced through the retained mixture coefficient, and any gain must come from the retained dirty score law. The diagnostic $L_{\mathrm{mix},+}$ captures this trade-off. Numerical certification requires componentwise bounds (Proposition 11) or direct audit of the selected set (Proposition 12). The threshold is fixed before the final contaminated calibration sample is used. Same-sample selection is handled only through a finite-grid empirical-CDF bound. The guarantees are marginal, not subgroup-conditional. Sharpness holds only within the scalar-discrepancy information class. Sharper bounds may exploit structure in P_{keep} or Q_{keep}, smoothness, or audit data.

Acknowledgments and Disclosure of Funding

The author gratefully acknowledges Research England E3 for funding the "MARS: Mathematical for AI in Real-world Systems" programme at Lancaster University. The author was supported by the China Scholarship Council under Grant Number No. 202208890004. The author is grateful to Chris Oates for helpful discussions and feedback on this work.

References

Anastasios Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Liviu Aolaritei, Zheyu Oliver Wang, Julie Zhu, Michael I Jordan, and Youssef Marzouk. Conformal prediction under lévy-prokhorov distribution shifts: Robustness to local and global perturbations. *arXiv preprint arXiv:2502.14105*, 2025.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Meshi Bashari, Matteo Sesia, and Yaniv Romano. Robust conformal outlier detection under contaminated reference data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 3091–3141. PMLR, 2025.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pages 2337–2363. PMLR, 2023.

Teresa Bortolotti, YX Wang, Xin Tong, Alessandra Menafoglio, Simone Vantini, and Matteo Sesia. Noise-adaptive conformal classification with marginal coverage. *arXiv preprint arXiv:2501.18060*, 2025.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, 119 (548):3033–3044, 2024a.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Predictive inference with weak supervision. *Journal of Machine Learning Research*, 25(118):1–45, 2024b.

Jase Clarkson, Wenkai Xu, Mihai Cucuringu, and Gesine Reinert. Split conformal prediction under data contamination. In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 230 of *Proceedings of Machine Learning Research*, pages 5–27. PMLR, 2024.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf`.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

Bat-Sheva Einbinder, Shai Feldman, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Label noise robustness of conformal prediction. *Journal of Machine Learning Research*, 25(328):1–66, 2024.

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

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

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.

Jackson Gorham and Lester Mackey. Measuring sample quality with stein's method. *Advances in neural information processing systems*, 28, 2015.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

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

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

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.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 3 1964.

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

Wittawat Jitkrittum, Heishiro Kanagawa, and Bernhard Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Conference on uncertainty in artificial intelligence*, pages 221–230. PMLR, 2020.

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.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

Qiang Liu, Lester Mackey, and Chris Oates. Probabilistic inference and learning with stein's method. *arXiv preprint arXiv:2603.07467*, 2026.

Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157–183, 2024.

Pascal Massart. The tight constant in the dvoretzky–kiefer–wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Chris J Oates. Minimum kernel discrepancy estimators. In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 133–161. Springer, 2022.

Roberto I Oliveira, Paulo Orenstein, Thiago Ramos, and Joao Vitor Romano. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(225):1–38, 2024.

Qian Peng, Yajie Bao, Haojie Ren, Zhaojun Wang, and Changliang Zou. Conformal prediction with cellwise outliers: A detect-then-impute approach. *arXiv preprint arXiv:2505.04986*, 2025.

Coby Penso and Jacob Goldberger. A conformal prediction score that is robust to label noise. *arXiv preprint arXiv:2405.02648*, 2024.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in artificial intelligence*, pages 844–853. PMLR, 2021.

Hongxiang Qiu, Edgar Dobriban, and Eric Tchetgen Tchetgen. Prediction sets adaptive to unknown covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5): 1680–1705, 2023.

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

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Matteo Sesia, YX Rachel Wang, and Xin Tong. Adaptive conformal classification with noisy labels. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):796–815, 2025.

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

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

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

Congye Wang, Ye Chen, Heishiro Kanagawa, and Chris J Oates. Stein Π-Importance Sampling. *Advances in Neural Information Processing Systems*, 36:71948–71994, 2023.

Huajun Xi, Kangdao Liu, Hao Zeng, Wenguang Sun, and Hongxin Wei. Exploring the noise robustness of online conformal prediction. *arXiv preprint arXiv:2501.18363*, 2025.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):943–965, 2024.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International conference on machine learning*, pages 25834–25866. PMLR, 2022.

A Appendix roadmap

Appendix B gives the fixed-threshold trimmed split-conformal algorithm. Appendix C collects supplementary generic theory, including independent threshold tuning, upper-coverage diagnostics, efficiency bounds, and marginalisation of high-probability conditional guarantees. Appendix D records quantile and edge-case conventions used in the finite-sample identity. Appendix E contains auxiliary diagnostic lemmas used in the retained-law and certificate analyses. Appendix F gives the proofs of the main results, including retained-law transfer, the exact finite-sample coverage identity, finite-sample scalar sharpness, same-sample threshold selection, and audit certificates. Appendix G presents benchmark regimes illustrating ideal separation and no-separation behaviour. Appendix H reports detailed experimental settings and additional simulations. Appendix I collects additional reporting conventions, diagnostic table variants, and guidance on interpreting empirical performance, population diagnostics, and conservative certificate bounds.

B Algorithm

For completeness, Algorithm 1 gives the fixed-threshold trimmed split-conformal procedure analysed in Section 3.1. The construction is standard split conformal prediction [Vovk et al., 2005, Lei et al., 2018] applied to the retained calibration subset

$$\{Z_i : S(Z_i) \le t^\star\}.$$

We use the conservative convention $\hat{\tau}_\alpha = +\infty$ when fewer than $\lceil 1/\alpha \rceil - 1$ points are retained, or when the conformal rank exceeds the retained sample size.

Algorithm 1 Fixed-threshold trimmed split conformal prediction

Require: Nonconformity score A, anomaly score S, calibration-independent threshold $t^\star$, contaminated calibration sample $\{Z_i = (X_i, Y_i)\}_{i=1}^m$, miscoverage level $\alpha \in (0, 1)$.
Ensure: Prediction set $C_\alpha(x)$.

for $i = 1, \ldots, m$ **do**
  Compute $A_i := A(Z_i)$ and $S_i := S(Z_i)$.
  Set $K_i := \mathbf{1}\{S_i \le t^\star\}$.
end for
Let $\mathcal{I}_{\text{keep}} := \{i : K_i = 1\}$ and $N_{\text{keep}} := |\mathcal{I}_{\text{keep}}|$.
if $N_{\text{keep}} = 0$ **then**
  Set $\hat{\tau}_\alpha := +\infty$.
else
  Set $r_{\text{keep}} := \lceil (N_{\text{keep}} + 1)(1 - \alpha) \rceil$.
  if $r_{\text{keep}} = N_{\text{keep}} + 1$ **then**
    Set $\hat{\tau}_\alpha := +\infty$.
  else
    Set $\hat{\tau}_\alpha$ to the r_{keep}-th order statistic of $\{A_i : i \in \mathcal{I}_{\text{keep}}\}$.
  end if
end if
Return $C_\alpha(x) := \{y \in \mathcal{Y} : A(x, y) \le \hat{\tau}_\alpha\}$.

C Supplementary generic theory

This appendix collects generic statements for implementation, diagnostics, and threshold tuning. They are not needed for the main retained-law transfer argument. Section 3 contains the fixed-threshold coverage theorem, the finite-sample scalar sharpness result, the same-sample finite-grid selection bound, and the audit certificates. This appendix keeps independent-tuning and auxiliary diagnostic material separate.

C.1 Unified notation for score-CDF gaps and mixture diagnostics

Unless stated otherwise, all quantities in this table are evaluated conditionally on the auxiliary sigma-field. Under this conditioning, A, S, and the threshold are fixed. Suprema are taken over $a \in \mathbb{R}$, and $F_M^A(a) := M\{A(Z) \le a\}$ denotes the score CDF under law M.

For the fixed threshold $t^\star$, write

$$K = \mathbf{1}\{S(Z) \le t^\star\}, \qquad p_c = P(K = 1), \qquad p_d = Q(K = 1),$$

and

$$\mu_{\mathrm{keep}} = (1 - \varepsilon)p_c + \varepsilon p_d.$$

When $p_d = 0$, Q_{keep} may be chosen arbitrarily, and all zero-weighted dirty terms are interpreted as zero. Componentwise clean-side quantities involving P_{keep} require $p_c > 0$.

Symbol	Definition	Role / condition
$Q_M^A(u)$	$\inf\{a \in \mathbb{R} : F_M^A(a) \ge u\}, u \in (0,1)$	Lower generalised score quantile
$P_{\mathrm{keep}}, Q_{\mathrm{keep}}$	$P(\cdot \mid S \le t^\star), Q(\cdot \mid S \le t^\star)$	Defined when the corresponding retention probability is positive
R	$P^\star(\cdot \mid S \le t^\star)$	Retained calibration law, requires $\mu_{\mathrm{keep}} > 0$
$\tilde{\varepsilon}_\star$	$\dfrac{\varepsilon p_d}{(1-\varepsilon)p_c + \varepsilon p_d}$	Retained mixture coefficient
$d^A_{M \to N,+}$	$\sup_a \big(F_M^A(a) - F_N^A(a)\big)_+$	One-sided score-CDF gap from M to N
$d^A_{R \to P,+}$	$\sup_a \big(F_R^A(a) - F_P^A(a)\big)_+$	Scalar loss controlling clean undercoverage
$d^A_{P \to R,+}$	$\sup_a \big(F_P^A(a) - F_R^A(a)\big)_+$	Mirror loss for upper-coverage diagnostics
$\Delta^A_{\mathrm{trim},+}$	$d^A_{P_{\mathrm{keep}} \to P,+}$	Clean-side trimming distortion, requires $p_c > 0$
$\Delta^A_{\mathrm{trim},-}$	$d^A_{P \to P_{\mathrm{keep}},+}$	Mirror clean-side distortion
Δ^A_{trim}	$\sup_a \lvert F^A_{P_{\mathrm{keep}}}(a) - F_P^A(a)\rvert = \max\{\Delta^A_{\mathrm{trim},+}, \Delta^A_{\mathrm{trim},-}\}$	Two-sided clean distortion
$D_{Q,+}$	$d^A_{Q_{\mathrm{keep}} \to P,+}$	Retained dirty discrepancy relevant to undercoverage
$D_{Q,-}$	$d^A_{P \to Q_{\mathrm{keep}},+}$	Mirror retained dirty discrepancy
D_Q	$\sup_a \lvert F^A_{Q_{\mathrm{keep}}}(a) - F_P^A(a)\rvert = \max\{D_{Q,+}, D_{Q,-}\}$	Two-sided retained dirty discrepancy
$\delta^{\mathrm{ref}}_{\mathrm{mix},+}$	$(1-\tilde{\varepsilon}_\star)\Delta^A_{\mathrm{trim},+} + \tilde{\varepsilon}_\star D_{Q,+}$	Componentwise upper bound on $d^A_{R \to P,+}$
$\delta^{\mathrm{ref}}_{\mathrm{mix},-}$	$(1-\tilde{\varepsilon}_\star)\Delta^A_{\mathrm{trim},-} + \tilde{\varepsilon}_\star D_{Q,-}$	Componentwise upper bound on $d^A_{P \to R,+}$
$\delta^{\mathrm{ref}}_{\mathrm{mix}}(D_Q)$	$(1-\tilde{\varepsilon}_\star)\Delta^A_{\mathrm{trim}} + \tilde{\varepsilon}_\star D_Q$	Symmetric fallback diagnostic
$\delta^{\mathrm{wc}}_{\mathrm{mix}}$	$(1-\tilde{\varepsilon}_\star)\Delta^A_{\mathrm{trim}} + \tilde{\varepsilon}_\star$	Worst-case dirty-side fallback, using $D_Q \le 1$
$L_{\mathrm{mix},+}$	$[1-\alpha-\delta^{\mathrm{ref}}_{\mathrm{mix},+}]_+$ unless a finite-sample L_{fs} version is explicitly stated	Population lower-bound diagnostic, not by itself a data-only certificate

Table 4: Unified notation for one-sided, mirror, two-sided, and mixture score-CDF diagnostics used in the appendix.

The notation $\delta^{\mathrm{ref}}_{\mathrm{mix},+}$ and $\delta^{\mathrm{ref}}_{\mathrm{mix},-}$ abbreviates the displayed quantities with inputs $D_{Q,+}$ and $D_{Q,-}$, respectively. The inequalities

$$d^A_{R \to P,+} \le \delta^{\mathrm{ref}}_{\mathrm{mix},+}, \qquad d^A_{P \to R,+} \le \delta^{\mathrm{ref}}_{\mathrm{mix},-}, \qquad \max\{d^A_{R \to P,+}, d^A_{P \to R,+}\} \le \delta^{\mathrm{ref}}_{\mathrm{mix}}(D_Q)$$

are upper bounds from the retained-law mixture representation and the triangle inequality. They are not exact decompositions in general.

C.2 Adaptive thresholds chosen on an independent tuning split

The theorem allows data-adaptive threshold selection without extra selection penalties, provided the tuning data are independent of the contaminated calibration sample used for the final conformal quantile. Same-sample threshold selection is handled separately in Proposition 10 and incurs an empirical-CDF uniformity penalty.

Theorem 3 (Independent adaptive threshold)**.** *Let $\mathcal{H}$ be an auxiliary or tuning sigma-field such that, conditionally on $\mathcal{G} \vee \mathcal{H}$, the contaminated calibration sample remains i.i.d. from $P^\star$ and independent of the clean test point. Suppose A, S, and a random threshold $\widehat{t}$ are $\mathcal{G} \vee \mathcal{H}$-measurable, and apply the trimmed split conformal procedure with $t^\star = \widehat{t}$. Define $p_c(\widehat{t})$, $p_d(\widehat{t})$, $\mu_{\mathrm{keep}}(\widehat{t})$, $R_{\widehat{t}}$, and $d^A_{R_{\widehat{t}} \to P,+}$ by replacing $t^\star$ with $\widehat{t}$ in Section 3.1. On the event $\mu_{\mathrm{keep}}(\widehat{t}) > 0$,*

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C^{\widehat{t}}_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G} \vee \mathcal{H}\big) \geq \big[1 - \alpha - d^A_{R_{\widehat{t}} \to P,+}\big]_+. \tag{9}$$

The exact identity (4) *also holds conditionally on $\mathcal{G} \vee \mathcal{H}$. All retained-law quantities are evaluated at $\widehat{t}$. As in Theorem 1, the identity uses generalised retained-score quantiles and does not require continuity of $F^A_{R_{\widehat{t}}}$.*

Proposition 5 (Oracle inequality for independent grid tuning)**.** *Let $\mathcal{T}$ be a finite set of candidate thresholds. For each $t \in \mathcal{T}$, let $B(t)$ be any population loss satisfying*

$$d^A_{R_t \to P,+} \leq B(t),$$

and let $W(t)$ be any population efficiency functional. Suppose an independent tuning stage produces $\widehat{B}(t)$ and $\widehat{W}(t)$, measurable with respect to $\mathcal{H}$, and that on an event $\mathcal{E}_{\mathrm{tune}} \in \mathcal{H}$,

$$\sup_{t \in \mathcal{T}} |\widehat{B}(t) - B(t)| \leq r_B, \qquad \sup_{t \in \mathcal{T}} |\widehat{W}(t) - W(t)| \leq r_W.$$

For $\eta \geq 0$, define

$$\widehat{\mathcal{T}}_\eta = \{t \in \mathcal{T} : \widehat{B}(t) \leq \eta\}.$$

Fix a calibration-independent fallback threshold t_{fb} with positive retained mass, for example $+\infty$ for no trimming. Define

$$\widehat{t} = \begin{cases} \arg\min_{t \in \widehat{\mathcal{T}}_\eta} \widehat{W}(t), & \widehat{\mathcal{T}}_\eta \neq \varnothing, \\ t_{\mathrm{fb}}, & \widehat{\mathcal{T}}_\eta = \varnothing, \end{cases}$$

with any deterministic tie-breaking rule. On $\mathcal{E}_{\mathrm{tune}} \cap \{\widehat{\mathcal{T}}_\eta \neq \varnothing\}$,

$$B(\widehat{t}) \leq \eta + r_B, \tag{10}$$

and hence

$$\mathbb{P}\{Y_{\mathrm{new}} \in C^{\widehat{t}}_\alpha(X_{\mathrm{new}}) \mid \mathcal{G} \vee \mathcal{H}\} \geq \big[1 - \alpha - \eta - r_B\big]_+.$$

If the oracle-feasible set $\{t \in \mathcal{T} : B(t) \leq \eta - r_B\}$ is nonempty, then on the same event

$$W(\widehat{t}) \leq \inf_{t \in \mathcal{T} : B(t) \leq \eta - r_B} W(t) + 2r_W. \tag{11}$$

Consequently, if

$$\mathbb{P}\big(\mathcal{E}_{\mathrm{tune}} \cap \{\widehat{\mathcal{T}}_\eta \neq \varnothing\}\big) \geq 1 - \beta - \beta_\varnothing,$$

then the marginal clean-test coverage is at least

$$[1 - \beta - \beta_\varnothing]_+ [1 - \alpha - \eta - r_B]_+ \geq 1 - \alpha - \eta - r_B - \beta - \beta_\varnothing,$$

provided the selected threshold has positive retained mass on the event where the tuning rule is used. If a certified fallback loss $B(t_{\mathrm{fb}}) \leq B_{\mathrm{fb}}$ is available, the bound can be sharpened. On the empty-feasible-set event, one can use the conditional fallback loss instead of subtracting $\beta_\varnothing$.

C.3 Operational certificate details

Proposition 11, stated and proved in Appendix F.9, clarifies how to use the population mixture upper-bound diagnostic. The proposition is deliberately modular. An implementation must provide a lower bound on clean retention, an upper bound on dirty retention, a clean one-sided distortion certificate, a retained dirty-side one-sided discrepancy certificate, and either the true contamination fraction or a defensible upper bound.

The worst-case dirty-side choice $B_{Q,+} = 1$ is always valid, but may be conservative. Proposition 12 gives a complementary route. If independent clean labelled audit data are available, the final selected set can be certified directly by exact binomial inversion. Selected score CDFs can also be certified by a one-sided KS band, without separately auditing every mixture component.

Audit sample size and certification margin. Proposition 12 provides an exact one sided binomial certificate for the coverage of a fixed final prediction set. Let M of n_{aud} independent clean audit observations be covered. At level $1 - \beta$, the lower confidence bound is

$$L_{\mathrm{bin}} = \mathrm{BetaInv}\{\beta, M, n_{\mathrm{aud}} - M + 1\}.$$

The final prediction set is certified to have clean coverage of at least c whenever $L_{\mathrm{bin}} \geq c$.

For example, at 95% confidence, 29 successes out of 29 observations yield a lower bound of approximately 0.902. This example also illustrates a limitation of auditing. If the true coverage is only slightly above the nominal target, a large audit sample may be needed to certify that target with high probability.

Auditing does not modify the prediction set. Its role is different from direct clean calibration, which constructs a new cut-off using clean observations.

Example: Non identifiability of clean coverage. Let the contaminated score distribution be $M^\star = \mathrm{Bernoulli}(1/2)$, with a contamination level $\varepsilon = 0.2$.

This observed distribution has two distinct Huber decompositions

$$P_1^A = Q_1^A = \mathrm{Bernoulli}(1/2),$$

and

$$P_2^A = \mathrm{Bernoulli}(5/8), \qquad Q_2^A = \delta_0.$$

Each decomposition satisfies

$$M^\star = (1 - \varepsilon)P_i^A + \varepsilon Q_i^A.$$

Yet the clean score CDFs differ at zero

$$F_{P_1}^A(0) = \frac{1}{2}, \qquad F_{P_2}^A(0) = \frac{3}{8}.$$

Hence, the same contaminated score distribution may correspond to different clean coverage probabilities at the same cut-off. Clean coverage is therefore not identifiable from the contaminated score distribution alone.

The additional information required by Proposition 11 expresses this identification limit. It is not an estimator missing from the trimming procedure. Independent clean audit data or structural assumptions are required to obtain a numerical certificate.

Remark 1 (Population and data-driven status of the certificates). The refined quantities in Theorem 1 are population quantities for the realised auxiliary stage. In simulations, they can be computed from the known data-generating mechanism and used as diagnostics. In applications, they become coverage certificates only after the bounds in Proposition 11 are justified, without reusing the final

contaminated calibration scores for post-hoc tuning. Alternatively, an independent audit sample can certify the final set through Proposition 12. If the certificate inputs hold only on a high-probability auxiliary event, Corollary 1 gives the corresponding marginal coverage reduction.

Remark 2 (Gaussian Mahalanobis certificate example). Suppose the clean covariates are certified to satisfy $X \sim N(\mu, \Sigma)$. Take

$$S(x) = (x-\mu)^\top \Sigma^{-1}(x-\mu),$$

with threshold $t = \chi^2_{d,q}$. Then the clean-retention certificate is $L_c = q$. If μ, Σ, and the threshold are estimated from an independent clean reference split, one may instead use a DKW-lowered empirical version [Dvoretzky et al., 1956, Massart, 1990].

Dirty retention and retained dirty-score discrepancy still require external information. One may assume or audit $p_d \le U_d$. When no retained-dirty score-law information is available, $B_{Q,+} = 1$ is the always-valid fallback. With these inputs, Proposition 11 gives a fully numerical clean-coverage lower bound. This example shows that the certificate interface is operational only when its inputs are independently justified.

C.4 Upper-bound and threshold-efficiency statements

Proposition 6 (One-sided coverage upper bound under no ties or randomised ties)**.** *Under the setup of Section 3.1, let*

$$R := P^\star(\cdot \mid S(Z) \le t^\star)$$

be the retained mixture law, and assume $\mu_{\mathrm{keep}} > 0$. The first inequality below requires only this retained-law positivity. The subsequent bounds involving $\delta^{\mathrm{ref}}_{\mathrm{mix},-}$ and $\delta^{\mathrm{ref}}_{\mathrm{mix}}(D_Q)$ additionally require $p_c > 0$.

If the distribution of $A(W)$, $W \sim R$, is continuous, then the ordinary inclusive trimmed split-conformal set satisfies

$$\begin{aligned}\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) &\le 1-\alpha + d^A_{P\to R,+} + \beta_m(\mu_{\mathrm{keep}}) \\ &\le 1-\alpha + \delta^{\mathrm{ref}}_{\mathrm{mix},-} + \beta_m(\mu_{\mathrm{keep}}) \\ &\le 1-\alpha + \delta^{\mathrm{ref}}_{\mathrm{mix}}(D_Q) + \beta_m(\mu_{\mathrm{keep}}). \end{aligned} \tag{12}$$

The same numerical upper bound holds without the continuity assumption under the following randomised lexicographic rule. Attach independent variables $U_i, U_0 \sim \mathrm{Unif}(0,1)$ to the retained calibration scores and the test score. Order the retained calibration pairs $(A(Z_i), U_i)$ lexicographically. Set $\widehat{T}_\alpha$ to the r_{keep}-th retained pair when $r_{\mathrm{keep}} \le N_{\mathrm{keep}}$, and to $(+\infty, 1)$ otherwise. Replace the prediction event by

$$\{(A(Z_{\mathrm{new}}), U_0) \le_{\mathrm{lex}} \widehat{T}_\alpha\}.$$

Here

$$\beta_m(\mu) := \mathbb{E}\left[\frac{1}{N+1}\right] = \frac{1-(1-\mu)^{m+1}}{(m+1)\mu}, \qquad N \sim \mathrm{Binomial}(m,\mu). \tag{13}$$

Thus, the coverage excess over $1-\alpha$ is controlled by the mirror one-sided retained-to-clean distortion, plus the usual finite-sample conformal granularity term. The final symmetric inequality is a conservative fallback.

The next proposition is an efficiency diagnostic, not a coverage guarantee. The coverage theorem above does not require its local density condition.

Proposition 7 (Threshold and width perturbation under a local density condition)**.** *Assume $\mu_{\mathrm{keep}} > 0$ and $p_c > 0$, so that R, P_{keep}, and $\delta^{\mathrm{ref}}_{\mathrm{mix}}(D_Q)$ are well defined. Let $p := 1-\alpha$, and let*

$$q_P(p) := \inf\{x \in \mathbb{R} : F^A_P(x) \ge p\}.$$

Assume $q_P(p) \in \mathbb{R}$, F^A_P is continuous at $q_P(p)$, and there exist $\lambda_{\min} > 0$ and $\rho > 0$ such that the two-sided CDF increment condition (28) *of Lemma 3 holds for $F = F^A_P$. A sufficient condition is that F^A_P is absolutely continuous and admits a density version satisfying $f^A_P(x) \ge \lambda_{\min}$ for Lebesgue-a.e. $x \in [q_P(p)-\rho, q_P(p)+\rho]$. Fix an integer*

$$n_0 \ge \lceil 1/\alpha \rceil - 1$$

and $\beta \in (0,1)$. Define

$$\eta_{n_0,\beta} := \sqrt{\frac{\log(2/\beta)}{2n_0}}, \qquad \Gamma_{n_0,\beta} := \delta^{\mathrm{ref}}_{\mathrm{mix}}(D_Q) + \eta_{n_0,\beta} + \frac{2}{n_0}.$$

If $\Gamma_{n_0,\beta} < \lambda_{\min}\rho$, then, conditional on the fixed auxiliary stage,

$$\mathbb{P}\left(|\hat{\tau}_\alpha - q_P(1-\alpha)| \le \frac{\Gamma_{n_0,\beta}}{\lambda_{\min}}\right) \ge 1 - \mathbb{P}(N_{\mathrm{keep}} < n_0) - \beta. \tag{14}$$

If, in addition, the nonconformity score is the absolute residual $A(x,y) = |y - \hat{f}(x)|$, so that retained scores are nonnegative and

$$C_\alpha(x) = [\hat{f}(x) - \hat{\tau}_\alpha, \hat{f}(x) + \hat{\tau}_\alpha],$$

then the same event implies

$$\mathrm{width}(C_\alpha(x)) \le 2q_P(1-\alpha) + \frac{2\Gamma_{n_0,\beta}}{\lambda_{\min}}.$$

Remark 3 (Role of the density condition). Proposition 7 is not needed for validity. It is an efficiency statement, so it requires additional regularity. In particular, the local density lower bound excludes atoms and flat regions near the clean oracle cutoff. For discrete or heavily tied nonconformity scores, the lower-bound theorem remains valid. However, this quantile-perturbation route to width control requires a separate discrete-score analysis.

C.5 Marginalising high-probability conditional guarantees

Corollary 1 (From conditional eventwise guarantees to marginal coverage)**.** *Let $\mathcal{G}$ be an auxiliary sigma-field, and let $\mathcal{E} \in \mathcal{G}$ be an event with*

$$\mathbb{P}(\mathcal{E}^c) \le \beta.$$

Suppose that, on $\mathcal{E}$,

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) \ge 1 - \alpha - L$$

for a nonnegative $\mathcal{G}$-measurable loss L, and that $L \le L_0$ on $\mathcal{E}$. Then

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}})\big) \ge (1-\beta)[1 - \alpha - L_0]_+. \tag{15}$$

In particular, since coverage probabilities are bounded by one,

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}})\big) \ge 1 - \alpha - L_0 - \beta. \tag{16}$$

For example, if

$$\mathcal{E} = \mathcal{E}_s \cap \mathcal{E}_{\mathrm{ref}}(\delta_\ell), \qquad \mathbb{P}(\mathcal{E}_s^c) \le \beta_s,$$

and the clean-reference DKW event [Dvoretzky et al., 1956, Massart, 1990] satisfies

$$\mathbb{P}\big(\mathcal{E}_{\mathrm{ref}}(\delta_\ell)^c\big) \le 2e^{-2\ell\delta_\ell^2},$$

then one may take

$$\beta = \beta_s + 2e^{-2\ell\delta_\ell^2}.$$

C.6 Contaminated-mixture evaluation

Proposition 8 (Coverage under the contaminated test distribution)**.** *Under the calibration and procedure setup of Section 3.1, suppose the test point is distributed as*

$$Z_{\mathrm{new}} \sim P^\star = (1-\varepsilon)P + \varepsilon Q$$

independently of the calibration sample, after conditioning on $\mathcal{G}$. Assume $\mu_{\mathrm{keep}} > 0$, and define

$$d^A_{R \to P^\star,+} := \sup_t \big(F^A_R(t) - F^A_{P^\star}(t)\big)_+.$$

Then, almost surely in $\mathcal{G}$,

$$\mathbb{P}_{Z_{\mathrm{new}} \sim P^\star}\{Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \mid \mathcal{G}\} \ge \big[1 - \alpha - d^A_{R \to P^\star,+}\big]_+. \tag{17}$$

The exact identity (4) *also holds with P replaced by $P^\star$. In particular, when $t^\star = +\infty$, $R = P^\star$, and the direct bound recovers the ordinary split-conformal lower bound $1 - \alpha$.*

The conservative clean-component reduction always remains valid

$$\mathbb{P}_{Z_{\mathrm{new}} \sim P^\star}\{Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \mid \mathcal{G}\} \geq (1-\varepsilon)\mathbb{P}_{Z_{\mathrm{new}} \sim P}\{Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \mid \mathcal{G}\}. \tag{18}$$

If, in addition, $p_c > 0$, applying the clean-target mixture upper-bound diagnostic gives

$$\mathbb{P}_{Z_{\mathrm{new}} \sim P^\star}\{Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \mid \mathcal{G}\} \geq (1-\varepsilon)\big[1 - \alpha - \delta^{\mathrm{ref}}_{\mathrm{mix}}(D_Q)\big]_+, \tag{19}$$

and the worst-case dirty-side version is

$$\mathbb{P}_{Z_{\mathrm{new}} \sim P^\star}\{Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \mid \mathcal{G}\} \geq (1-\varepsilon)\big[1 - \alpha - \delta^{\mathrm{wc}}_{\mathrm{mix}}\big]_+. \tag{20}$$

The last two bounds are deliberately conservative, because they ignore any coverage of the dirty test component.

Remark 4 (Choosing the fixed threshold). Throughout the theoretical analysis, $t^\star$ is treated as fixed after conditioning on the clean auxiliary stage. This is essential for the pointwise-thinning argument in Theorem 1. Once the auxiliary clean data have been realised, the retention indicators

$$K_i = \mathbf{1}\{S_i \leq t^\star\}$$

must be defined by applying a deterministic threshold separately to each calibration observation.

In practice, however, $t^\star$ need not be chosen arbitrarily. Several principled strategies are available. One can choose $t^\star$ using external validation data or prior domain knowledge about anomalous values of S. Another option is a sensitivity analysis over a prespecified threshold grid, reporting coverage and retained sample size across the grid.

A third option is to tune $t^\star$ using an auxiliary clean reference set, when available. More generally, Theorem 3 allows arbitrary threshold-selection rules based on an independent tuning split. Proposition 5 gives a finite-grid oracle inequality when the tuning split provides uniform estimates of coverage loss and efficiency.

In implementations with a fitted auxiliary anomaly score, the clean reference split is used only for threshold selection. The anomaly score itself is already fixed after the auxiliary score-fitting stage.

Our theory makes clear that the choice of $t^\star$ affects robustness through two distinct quantities:

$$\tilde{\varepsilon}_\star = \frac{\varepsilon p_d}{(1-\varepsilon)p_c + \varepsilon p_d}, \qquad \Delta^A_{\mathrm{trim}} = \sup_{t \in \mathbb{R}} \big| F^A_{P_{\mathrm{keep}}}(t) - F^A_P(t) \big|.$$

The first term, $\tilde{\varepsilon}_\star$, is the retained mixture coefficient of the dirty component in the retained calibration set. The second term, Δ^A_{trim}, measures how much the same threshold distorts the clean nonconformity distribution.

This suggests a practical rule for choosing $t^\star$. If the main goal is reliable clean-test coverage, one should first keep Δ^A_{trim} small. Overly aggressive trimming can strongly perturb the clean conformal quantile, even when it reduces the retained mixture coefficient.

When S clearly separates clean and contaminated observations, more aggressive trimming can help. It can reduce $\tilde{\varepsilon}_\star$ without causing a comparable increase in Δ^A_{trim}. In weak- or no-separation regimes, aggressive trimming is unlikely to help. It mainly removes clean calibration points and increases clean trimming distortion, without substantially reducing the retained mixture coefficient.

If an auxiliary clean reference split is available, a natural choice is a clean-reference quantile rule for $t^\star$. This stabilises the clean retention probability at a target level. More broadly, threshold selection balances two objectives: making contaminated points unlikely to survive trimming, and keeping the retained clean nonconformity distribution close to its untrimmed version.

D Technical remarks and edge cases

Lemma 1 (Generalised quantiles commute with order statistics)**.** *Let $Q : (0,1) \to \mathbb{R} \cup \{\pm\infty\}$ be non-decreasing, and let $U_1, \ldots, U_n$ be real numbers in $(0,1)$. Set $X_i = Q(U_i)$. Then for every $r \in \{1, \ldots, n\}$,*

$$X_{(r)} = Q(U_{(r)}),$$

where order statistics are taken in non decreasing order. Consequently, if $Q = Q_F$ is the lower generalised quantile of a distribution function F, and $U_1, \dots, U_n \overset{\text{iid}}{\sim} \text{Unif}(0,1)$, then $Q_F(U_i) \sim F$. The r-th order statistic of the transformed sample has the same distribution as $Q_F(U_{(r)})$, even when F has atoms.

Proof. Sort the uniforms as $U_{(1)} \leq \cdots \leq U_{(n)}$. Since Q is non-decreasing, the transformed values satisfy

$$Q(U_{(1)}) \leq \cdots \leq Q(U_{(n)}),$$

with equalities allowed. Hence the r-th ordered transformed value is $Q(U_{(r)})$. The distributional statement follows from the generalised inverse transform. □

Remark 5 (Degenerate infinite-threshold event). Because the theorem gives a lower bound, it is important to isolate the event on which the algorithm returns the degenerate cutoff $\hat{\tau}_\alpha = +\infty$. Since the retention indicators

$$K_i = \mathbf{1}\{S(Z_i) \leq t^\star\}$$

are i.i.d. Bernoulli(μ_{keep}), we have

$$N_{\text{keep}} \sim \text{Binomial}\big(m, \mu_{\text{keep}}\big).$$

Define

$$n_{\infty,\alpha} := \max\Big\{0, \lceil \alpha^{-1} \rceil - 2\Big\}.$$

By the algorithmic definition of the cutoff,

$$\hat{\tau}_\alpha = +\infty \qquad \Longleftrightarrow \qquad N_{\text{keep}} \leq n_{\infty,\alpha}.$$

For the experimental nominal level $\alpha = 0.1$, this threshold is $n_{\infty,\alpha} = 8$. Retaining eight or fewer calibration points forces the uninformative infinite cutoff. Equivalently,

$$\mathbb{P}\big(\hat{\tau}_\alpha = +\infty\big) = \sum_{j=0}^{n_{\infty,\alpha}} \binom{m}{j} \mu_{\text{keep}}^j (1 - \mu_{\text{keep}})^{m-j}.$$

In particular,

$$\mathbb{P}\big(N_{\text{keep}} = 0\big) = (1 - \mu_{\text{keep}})^m \leq e^{-m\mu_{\text{keep}}}.$$

If $m\mu_{\text{keep}} > n_{\infty,\alpha}$, Hoeffding's inequality [Hoeffding, 1963] further gives

$$\mathbb{P}\big(\hat{\tau}_\alpha = +\infty\big) \leq \exp\left(-2m\left(\mu_{\text{keep}} - \frac{n_{\infty,\alpha}}{m}\right)^2\right).$$

This event is controlled explicitly by the retained mass μ_{keep}. It should be interpreted as an efficiency failure, not a validity failure. If $C_\alpha(x) = \mathcal{Y}$ on this event, then the conditional coverage is indeed one. However, the prediction set is uninformative. For symmetric interval-valued conformal rules, this corresponds to infinite width.

Remark 6 (Comparison with the ordinary untrimmed baseline). The theorem specialises immediately to ordinary split conformal prediction. Indeed, if one sets $t^\star = +\infty$, then every calibration point is retained, so

$$p_c = p_d = 1, \qquad P_{\text{keep}} = P, \qquad \Delta^A_{\text{trim}} = 0, \qquad \tilde{\varepsilon}_\star = \varepsilon.$$

In that case $Q_{\text{keep}} = Q$, and the retained-dirty discrepancy reduces to

$$D_Q^{\text{untrim}} := \sup_{t \in \mathbb{R}} \big| F_Q^A(t) - F_P^A(t) \big|.$$

Therefore Theorem 1 yields the theorem-level lower bound for the untrimmed rule

$$\mathbb{P}\big(Y_{\text{new}} \in C_\alpha^{\text{untrim}}(X_{\text{new}}) \,\big|\, \mathcal{G}\big) \geq 1 - \alpha - \varepsilon D_Q^{\text{untrim}}.$$

Comparing this with the trimmed lower bound

$$1 - \alpha - (1 - \tilde{\varepsilon}_\star)\Delta^A_{\text{trim}} - \tilde{\varepsilon}_\star D_Q,$$

shows that the trimmed theorem-level lower bound is sharper than the untrimmed theorem-level lower bound whenever

$$(1-\tilde{\varepsilon}_\star)\Delta^A_{\mathrm{trim}} + \tilde{\varepsilon}_\star D_Q < \varepsilon D_Q^{\mathrm{untrim}}. \tag{21}$$

This comparison isolates the paper's central trade-off. Trimming improves the lower bound only when the reduction in retained contamination is large enough to offset the clean-law distortion caused by filtering. The comparison is intentionally one-sided. It concerns theorem-level lower bounds, and does not by itself imply actual coverage dominance, efficiency improvement, or width reduction.

Remark 7 (On the clean trimming distortion Δ^A_{trim}). The quantity

$$\Delta^A_{\mathrm{trim}} = \sup_{t\in\mathbb{R}} \left| F^A_{P_{\mathrm{keep}}}(t) - F^A_P(t) \right|$$

measures how much the distribution of the nonconformity score $A(Z)$ is distorted by the fixed-threshold trimming rule. Here F^A_P denotes the CDF of $A(Z)$ when $Z \sim P$, while $F^A_{P_{\mathrm{keep}}}$ denotes the CDF of $A(Z)$ under the retained clean law

$$P_{\mathrm{keep}} := P(\,\cdot \mid S(Z) \le t^\star),$$

that is, the conditional law of a clean point given retention by the population trimming rule induced by the anomaly score S and fixed threshold $t^\star$. The assumption $p_c > 0$ in Theorem 1 ensures that this conditional law, and hence Δ^A_{trim}, is well-defined.

This quantity is not merely a distributional discrepancy. Split conformal prediction is driven by a high quantile of the clean nonconformity law. Therefore, changing the retained-clean distribution directly perturbs the conformal cutoff. In particular, Corollary 2 shows that, under a local density lower bound near the clean $(1-\alpha)$-quantile, the induced quantile shift is controlled linearly by Δ^A_{trim}. Thus, Δ^A_{trim} should be read as a direct measure of how much trimming can move the clean conformal threshold. It is not merely an abstract discrepancy between two distributions.

In particular, Δ^A_{trim} captures the dependence between the anomaly score S and the nonconformity score A under the clean distribution P

- If S and A are independent under P, then conditioning on the event $\{S(Z) \le t^\star\}$ does not change the marginal distribution of $A(Z)$. Consequently,

  $$F^A_{P_{\mathrm{keep}}} \equiv F^A_P, \qquad \Delta^A_{\mathrm{trim}} = 0.$$

  In this case, trimming does not distort the clean conformal score distribution. The loss term in the population coverage bound therefore comes from residual retained contamination. Separate finite-sample effects, such as the probability of a degenerate infinite threshold and the granularity of the retained conformal quantile, affect efficiency or certificate sharpness. They do not add a separate loss term to the finite-sample rank-valid retained-law conformal step.

- If S and A are positively associated under P, trimming may preferentially remove part of the right tail of the clean score distribution. For example, "hard" clean samples may both look more anomalous under S and have larger nonconformity scores A. In this case, the retained score distribution is typically shifted toward smaller values than the full clean score distribution. Equivalently, $F^A_{P_{\mathrm{keep}}}$ is typically larger than F^A_P at relevant thresholds, and $\Delta^A_{\mathrm{trim}} > 0$. This creates a potential under-coverage effect, because the retained calibration scores can be systematically smaller than the full clean calibration scores.

From a modelling perspective, Δ^A_{trim} is a sensitivity parameter. It quantifies how much the trimming rule changes the clean calibration distribution. Thus, when designing S and A, one should choose an anomaly score S that separates contaminated points well, while remaining as weakly dependent as possible on the clean nonconformity score A.

E Auxiliary and diagnostic results

E.1 Diagnostic retained-count concentration

The following concentration lemma is used as a diagnostic retained-count certificate and to control rare small-retained-sample events. It is not needed for the proof of Theorem 1, which conditions directly on the realised retained sample size.

Lemma 2 (Diagnostic retained-count and contamination bounds for fixed-threshold trimming)**.** *Let $Z_1, \dots, Z_m$ be generated under the ε-contamination model [Huber, 1964]*

$$P^\star := (1-\varepsilon)P + \varepsilon Q, \qquad 0 \le \varepsilon < 1.$$

For bookkeeping, realise this mixture on an expanded probability space through independent latent labels

$$C_i \sim \mathrm{Bernoulli}(1-\varepsilon), \qquad D_i := 1 - C_i,$$

with

$$Z_i \mid \{C_i = 1\} \sim P, \qquad Z_i \mid \{D_i = 1\} \sim Q,$$

independently across $i = 1, \dots, m$. Marginally, each Z_i has law $P^\star$.

Let $S : \mathcal{Z} \to \mathbb{R}$ be a measurable score function and fix a deterministic threshold $t^\star \in \overline{\mathbb{R}}$. For each $i = 1, \dots, m$, define

$$K_i := \mathbf{1}\{S(Z_i) \le t^\star\},$$

and define the retained clean, dirty, and total counts

$$N_{\mathrm{clean}} := \sum_{i=1}^m C_i K_i, \qquad N_{\mathrm{dirty}} := \sum_{i=1}^m D_i K_i, \qquad N_{\mathrm{keep}} := \sum_{i=1}^m K_i = N_{\mathrm{clean}} + N_{\mathrm{dirty}}.$$

Let

$$p_c := \mathbb{P}_{Z\sim P}\big(S(Z) \le t^\star\big), \qquad p_d := \mathbb{P}_{Z\sim Q}\big(S(Z) \le t^\star\big),$$

and set

$$\pi_c := (1-\varepsilon)p_c, \qquad \pi_d := \varepsilon p_d, \qquad \mu_{\mathrm{keep}} := \pi_c + \pi_d = (1-\varepsilon)p_c + \varepsilon p_d.$$

Then the following bounds hold.

*(i) (*Retained-count lower bound*) For any $\eta > 0$,*

$$\mathbb{P}\Big(N_{\mathrm{keep}} \ge m(\mu_{\mathrm{keep}} - \eta)\Big) \ge 1 - \exp(-2m\eta^2). \tag{22}$$

*(ii) (*Simultaneous clean, dirty, and retained count bounds*) For any $\eta > 0$, with probability at least*

$$1 - 3\exp(-2m\eta^2),$$

the following inequalities hold simultaneously

$$N_{\mathrm{clean}} \ge m(\pi_c - \eta), \tag{23}$$
$$N_{\mathrm{dirty}} \le m(\pi_d + \eta), \tag{24}$$
$$N_{\mathrm{keep}} \ge m(\mu_{\mathrm{keep}} - \eta). \tag{25}$$

The factor of three is the cost of a simple union bound. Since $N_{\mathrm{keep}} = N_{\mathrm{clean}} + N_{\mathrm{dirty}}$, sharper simultaneous multinomial concentration inequalities could improve this diagnostic constant. They are not needed for the retained-law validity theorem.

*(iii) (*Retained contamination proportion*) For any $\eta_d > 0$ and any $\eta_k \in (0, \mu_{\mathrm{keep}})$, with probability at least*

$$1 - \exp(-2m\eta_d^2) - \exp(-2m\eta_k^2),$$

one has

$$\frac{N_{\mathrm{dirty}}}{N_{\mathrm{keep}}} \le \frac{\pi_d + \eta_d}{\mu_{\mathrm{keep}} - \eta_k} = \frac{\varepsilon p_d + \eta_d}{(1-\varepsilon)p_c + \varepsilon p_d - \eta_k}. \tag{26}$$

In particular, taking $\eta_d = \eta_k = \eta \in (0, \mu_{\mathrm{keep}})$ gives the simpler bound

$$\frac{N_{\mathrm{dirty}}}{N_{\mathrm{keep}}} \le \frac{\pi_d + \eta}{\mu_{\mathrm{keep}} - \eta} = \frac{\varepsilon p_d + \eta}{(1-\varepsilon)p_c + \varepsilon p_d - \eta} \tag{27}$$

with probability at least $1 - 2\exp(-2m\eta^2)$.

Proof. Define

$$X_i^{(c)} := C_i K_i, \qquad X_i^{(d)} := D_i K_i, \qquad i = 1, \dots, m.$$

Each of $X_i^{(c)}$, $X_i^{(d)}$, and K_i takes values in $\{0, 1\}$. Since the latent labels and observations are independent across i, and the threshold $t^\star$ is fixed, each of the three sequences

$$\{X_i^{(c)}\}_{i=1}^m, \qquad \{X_i^{(d)}\}_{i=1}^m, \qquad \{K_i\}_{i=1}^m$$

is individually i.i.d. Bernoulli across i. Within a single index i, however, the variables are functionally related through

$$K_i = X_i^{(c)} + X_i^{(d)}.$$

Only cross-i independence is used below, through separate Hoeffding bounds [Hoeffding, 1963] and union bounds. By construction,

$$N_{\text{clean}} = \sum_{i=1}^m X_i^{(c)}, \qquad N_{\text{dirty}} = \sum_{i=1}^m X_i^{(d)}, \qquad N_{\text{keep}} = \sum_{i=1}^m K_i.$$

Under the latent-label realisation of the ε-contamination model,

$$\mathbb{P}(X_i^{(c)} = 1) = \mathbb{P}(C_i = 1,\ S(Z_i) \le t^\star) = (1-\varepsilon)\mathbb{P}_{Z\sim P}\big(S(Z) \le t^\star\big) = \pi_c.$$

Similarly,

$$\mathbb{P}(X_i^{(d)} = 1) = \varepsilon \mathbb{P}_{Z\sim Q}\big(S(Z) \le t^\star\big) = \pi_d,$$

and

$$\mathbb{P}(K_i = 1) = (1-\varepsilon)p_c + \varepsilon p_d = \mu_{\text{keep}}.$$

Hoeffding's one-sided lower-tail inequality [Hoeffding, 1963] for i.i.d. variables $Y_1, \dots, Y_m \in [0, 1]$ gives

$$\mathbb{P}\Big(\frac{1}{m}\sum_{i=1}^m Y_i - \mathbb{E}[Y_1] \le -\eta\Big) \le \exp(-2m\eta^2).$$

Applying this with $Y_i = K_i$ proves (22). Applying one-sided Hoeffding bounds to

$$\sum_{i=1}^m X_i^{(c)}, \qquad \sum_{i=1}^m X_i^{(d)}, \qquad \sum_{i=1}^m K_i,$$

and taking a union bound gives the following conservative simultaneous control. This uses the deterministic identity $K_i = X_i^{(c)} + X_i^{(d)}$ only for bookkeeping. A multinomial concentration argument could sharpen the constants.

$$\mathbb{P}\Big(N_{\text{clean}} \ge m(\pi_c - \eta),\ N_{\text{dirty}} \le m(\pi_d + \eta),\ N_{\text{keep}} \ge m(\mu_{\text{keep}} - \eta)\Big) \ge 1 - 3\exp(-2m\eta^2),$$

proving (23)–(25).

On the event

$$\Big\{N_{\text{dirty}} \le m(\pi_d + \eta_d)\Big\} \cap \Big\{N_{\text{keep}} \ge m(\mu_{\text{keep}} - \eta_k)\Big\},$$

the denominator is positive because $\eta_k < \mu_{\text{keep}}$, and

$$\frac{N_{\text{dirty}}}{N_{\text{keep}}} \le \frac{m(\pi_d + \eta_d)}{m(\mu_{\text{keep}} - \eta_k)} = \frac{\pi_d + \eta_d}{\mu_{\text{keep}} - \eta_k}.$$

Hoeffding's upper-tail bound for N_{dirty} and lower-tail bound for N_{keep}, followed by a union bound, give probability at least

$$1 - \exp(-2m\eta_d^2) - \exp(-2m\eta_k^2).$$

This proves (26). The one-parameter version is the special case $\eta_d = \eta_k$. □

E.2 Quantile perturbation and trimming-distortion identities

Lemma 3 (Quantile perturbation under Kolmogorov distance)**.** *Let F and G be cumulative distribution functions on $\mathbb{R}$. For $p \in (0,1)$, define the lower generalised p-quantiles*

$$q_F(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}, \qquad q_G(p) := \inf\{x \in \mathbb{R} : G(x) \geq p\}.$$

Let $x_0 = q_F(p)$. Assume $x_0 \in \mathbb{R}$, F is continuous at x_0, and there exist constants $\lambda_{\min} > 0$ and $\rho > 0$ such that, for every $h \in [0, \rho]$,

$$F(x_0 + h) - F(x_0) \geq \lambda_{\min} h, \qquad F(x_0) - F(x_0 - h) \geq \lambda_{\min} h. \tag{28}$$

Suppose also that

$$\sup_{t \in \mathbb{R}} |F(t) - G(t)| \leq \Delta \tag{29}$$

and

$$\Delta < \lambda_{\min} \rho. \tag{30}$$

Then

$$|q_G(p) - q_F(p)| \leq \frac{\Delta}{\lambda_{\min}}. \tag{31}$$

In particular, $q_G(p) \in [x_0 - \rho, x_0 + \rho]$. Condition (28) *holds, for example, if F is absolutely continuous and has a density version satisfying $f_F(x) \geq \lambda_{\min}$ for Lebesgue-a.e. $x \in [x_0 - \rho, x_0 + \rho]$.*

Proof. By the definition of $x_0 = q_F(p)$, $F(x) < p$ for all $x < x_0$, and $F(x_0) \geq p$. Continuity at x_0 gives $F(x_0) = p$.

Set $h_+ = \Delta / \lambda_{\min}$. By (30), $h_+ < \rho$ unless $\Delta = 0$, in which case $h_+ = 0 \leq \rho$. From (28),

$$F(x_0 + h_+) \geq p + \Delta.$$

Therefore (29) implies

$$G(x_0 + h_+) \geq F(x_0 + h_+) - \Delta \geq p,$$

so $q_G(p) \leq x_0 + \frac{\Delta}{\lambda_{\min}}$.

For the lower bound, let $\xi > 0$ and set $h_\xi = (\Delta + \xi)/\lambda_{\min}$. For sufficiently small ξ, $h_\xi \leq \rho$. Again by (28),

$$F(x_0 - h_\xi) \leq p - (\Delta + \xi).$$

Hence

$$G(x_0 - h_\xi) \leq F(x_0 - h_\xi) + \Delta \leq p - \xi < p.$$

By the definition of the lower generalised quantile,

$$q_G(p) \geq x_0 - h_\xi = x_0 - \frac{\Delta + \xi}{\lambda_{\min}}.$$

Letting $\xi \downarrow 0$ gives $q_G(p) \geq x_0 - \frac{\Delta}{\lambda_{\min}}$. Combining the two bounds proves (31). The final inclusion follows from (30). □

Remark 8 (Scope of the quantile perturbation lemma). Lemma 3 is an efficiency and stability tool, not a prerequisite for conformal validity. The strict inequality in (30) is intentional. At the boundary $\Delta = \lambda_{\min}\rho$, the local CDF-increment information may not extend far enough to support the left-tail perturbation argument without enlarging the neighbourhood. If the score law has atoms, or if the CDF is locally flat near $q_F(p)$, the lemma may not apply. Discrete nonconformity scores require a separate perturbation analysis.

Lemma 4 (Universal upper bound for the clean trimming distortion)**.** *Let P be the clean distribution, let $S : \mathcal{Z} \to \mathbb{R}$ be a measurable score, and fix a deterministic threshold $t^\star \in \overline{\mathbb{R}}$. Define*

$$p_c := \mathbb{P}_{Z \sim P}\big(S(Z) \leq t^\star\big).$$

Assume $p_c > 0$, and define the retained clean law

$$P_{\text{keep}} := P(\,\cdot \mid S(Z) \leq t^\star).$$

Let

$$\Delta^A_{\mathrm{trim}} := \sup_{t\in\mathbb{R}} \big|F^A_{P_{\mathrm{keep}}}(t) - F^A_P(t)\big|,$$

where $F^A_{P_{\mathrm{keep}}}$ and F^A_P denote the cumulative distribution functions of $A(Z)$ under $Z \sim P_{\mathrm{keep}}$ and $Z \sim P$, respectively.

The case $t^\star = +\infty$ is included. Then $p_c = 1$, $P_{\mathrm{keep}} = P$, and the bound below gives zero distortion. Then

$$\Delta^A_{\mathrm{trim}} \le \mathrm{TV}\big(P^A_{\mathrm{keep}}, P^A\big) \le \mathrm{TV}\big(P_{\mathrm{keep}}, P\big) = 1 - p_c. \tag{32}$$

Proof. By definition of Δ^A_{trim},

$$\Delta^A_{\mathrm{trim}} = \sup_{t\in\mathbb{R}} \big|P^A_{\mathrm{keep}}((-\infty, t]) - P^A((-\infty, t])\big|.$$

Since the half-lines $(-\infty, t]$ form a subclass of all measurable sets on $\mathbb{R}$, we immediately obtain

$$\Delta^A_{\mathrm{trim}} \le \sup_{B\in\mathcal{B}(\mathbb{R})} \big|P^A_{\mathrm{keep}}(B) - P^A(B)\big| = \mathrm{TV}\big(P^A_{\mathrm{keep}}, P^A\big).$$

Next, P^A_{keep} and P^A are pushforwards of P_{keep} and P under the measurable map A. For any Borel set $B \subseteq \mathbb{R}$,

$$P^A_{\mathrm{keep}}(B) - P^A(B) = P_{\mathrm{keep}}\big(A^{-1}(B)\big) - P\big(A^{-1}(B)\big),$$

hence

$$\mathrm{TV}\big(P^A_{\mathrm{keep}}, P^A\big) \le \mathrm{TV}\big(P_{\mathrm{keep}}, P\big).$$

It remains to compute $\mathrm{TV}(P_{\mathrm{keep}}, P)$. Since $P_{\mathrm{keep}} = P(\cdot \mid S(Z) \le t^\star)$, its Radon–Nikodym derivative with respect to P is

$$\frac{dP_{\mathrm{keep}}}{dP}(z) = \frac{\mathbf{1}\{S(z) \le t^\star\}}{p_c}.$$

Therefore,

$$\begin{aligned}
\mathrm{TV}\big(P_{\mathrm{keep}}, P\big) &= \frac{1}{2}\int \left|\frac{\mathbf{1}\{S(z) \le t^\star\}}{p_c} - 1\right| dP(z) \\
&= \frac{1}{2}\left[\int_{\{S\le t^\star\}} \left(\frac{1}{p_c} - 1\right) dP + \int_{\{S > t^\star\}} 1\, dP\right] \\
&= \frac{1}{2}\left[p_c\left(\frac{1}{p_c} - 1\right) + (1 - p_c)\right] \\
&= \frac{1}{2}\left[(1 - p_c) + (1 - p_c)\right] = 1 - p_c,
\end{aligned}$$

where we used

$$p_c\left(\frac{1}{p_c} - 1\right) = 1 - p_c.$$

Combining the three displays proves (32). □

Proposition 9 (Covariance and retained–discarded representations of the clean trimming distortion)**.** *Let P be the clean distribution, let $S : \mathcal{Z} \to \mathbb{R}$ be a measurable score, and fix a deterministic threshold $t^\star \in \mathbb{R}$. Let*

$$K := \mathbf{1}\{S(Z) \le t^\star\}, \qquad p_c := \mathbb{P}_{Z\sim P}(K = 1),$$

and assume $p_c > 0$. Define the retained clean law

$$P_{\mathrm{keep}} := P(\cdot \mid S(Z) \le t^\star),$$

and the clean trimming distortion

$$\Delta^A_{\mathrm{trim}} := \sup_{t\in\mathbb{R}} \big|F^A_{P_{\mathrm{keep}}}(t) - F^A_P(t)\big|.$$

Then for every $t \in \mathbb{R}$,

$$F^A_{P_{\mathrm{keep}}}(t) - F^A_P(t) = \frac{\mathrm{Cov}_P\big(\mathbf{1}\{A(Z) \le t\},\, K\big)}{p_c}. \tag{33}$$

Consequently,

$$\Delta^A_{\mathrm{trim}} = \frac{1}{p_c} \sup_{t\in\mathbb{R}} \big|\mathrm{Cov}_P\big(\mathbf{1}\{A(Z) \le t\},\, K\big)\big|. \tag{34}$$

If $p_c < 1$, define the discarded clean law

$$P_{\mathrm{drop}} := P(\,\cdot \mid S(Z) > t^\star).$$

Then the clean trimming distortion also admits the exact retained–discarded representation

$$\Delta^A_{\mathrm{trim}} = (1-p_c) \sup_{t\in\mathbb{R}} \big|F^A_{P_{\mathrm{keep}}}(t) - F^A_{P_{\mathrm{drop}}}(t)\big|. \tag{35}$$

If $p_c = 1$, then $P_{\mathrm{keep}} = P$ and $\Delta^A_{\mathrm{trim}} = 0$.

Proof. For any $t \in \mathbb{R}$,

$$F^A_{P_{\mathrm{keep}}}(t) = P\big(A(Z) \le t \mid K = 1\big) = \frac{\mathbb{E}_P[\mathbf{1}\{A(Z) \le t\}K]}{p_c}.$$

Therefore,

$$\begin{aligned} F^A_{P_{\mathrm{keep}}}(t) - F^A_P(t) &= \frac{\mathbb{E}_P[\mathbf{1}\{A(Z) \le t\}K]}{p_c} - \mathbb{E}_P[\mathbf{1}\{A(Z) \le t\}] \\ &= \frac{\mathbb{E}_P[\mathbf{1}\{A(Z) \le t\}K] - p_c\,\mathbb{E}_P[\mathbf{1}\{A(Z) \le t\}]}{p_c} \\ &= \frac{\mathrm{Cov}_P\big(\mathbf{1}\{A(Z) \le t\},\, K\big)}{p_c}, \end{aligned}$$

which proves (33). Taking the supremum over t yields (34).

If $p_c = 1$, then $K = 1$ P-almost surely, so $P_{\mathrm{keep}} = P$ and $\Delta^A_{\mathrm{trim}} = 0$.

It remains to consider the case $p_c < 1$. Define

$$P_{\mathrm{drop}} := P(\,\cdot \mid K = 0) = P(\,\cdot \mid S(Z) > t^\star).$$

Then, for every $t \in \mathbb{R}$,

$$F^A_P(t) = p_c F^A_{P_{\mathrm{keep}}}(t) + (1-p_c) F^A_{P_{\mathrm{drop}}}(t).$$

Therefore

$$\begin{aligned} F^A_{P_{\mathrm{keep}}}(t) - F^A_P(t) &= F^A_{P_{\mathrm{keep}}}(t) - \Big[p_c F^A_{P_{\mathrm{keep}}}(t) + (1-p_c) F^A_{P_{\mathrm{drop}}}(t)\Big] \\ &= (1-p_c)\Big(F^A_{P_{\mathrm{keep}}}(t) - F^A_{P_{\mathrm{drop}}}(t)\Big). \end{aligned}$$

Taking the supremum over $t \in \mathbb{R}$ proves (35). □

Remark 9 (Interpretation of the covariance and retained–discarded identities). The main content of Proposition 9 is structural. The covariance identity (33) and its supremum form (34) show that clean trimming distortion is driven by dependence under the clean law P. Specifically, it depends on the relation between the retention indicator $K = \mathbf{1}\{S(Z) \le t^\star\}$ and the nonconformity events $\{A(Z) \le t\}$.

When $p_c < 1$, the retained–discarded identity (35) gives an equivalent and often more interpretable form

$$\Delta^A_{\mathrm{trim}} = (1-p_c) d_{\mathrm{Kol}}\big(F^A_{P_{\mathrm{keep}}}, F^A_{P_{\mathrm{drop}}}\big),$$

where

$$d_{\mathrm{Kol}}(F, G) := \sup_{t\in\mathbb{R}} |F(t) - G(t)|.$$

Thus, clean trimming distortion is small when little clean mass is discarded ($1 - p_c \approx 0$), or when retained and discarded clean points have similar nonconformity-score distributions. Taking $d_{\mathrm{Kol}} \leq 1$ immediately gives the total-variation bound

$$\Delta^A_{\mathrm{trim}} \leq 1 - p_c.$$

When $p_c = 1$, this bound is trivial because $\Delta^A_{\mathrm{trim}} = 0$.

Corollary 2 (Quantile stability induced by trimming)**.** *Let F^A_P denote the cumulative distribution function of $A(Z)$ under $Z \sim P$, and let*

$$q_P(1-\alpha) := \inf\{x \in \mathbb{R} : F^A_P(x) \geq 1 - \alpha\}$$

be its lower generalised $(1-\alpha)$-quantile. Let

$$q_{\mathrm{keep}}(1-\alpha) := \inf\{x \in \mathbb{R} : F^A_{P_{\mathrm{keep}}}(x) \geq 1 - \alpha\}$$

be the corresponding retained-clean quantile.

Assume the setting of Lemma 4. Further assume that F^A_P is continuous at $q_P(1-\alpha)$. Also assume that there exist constants $\lambda_{\min} > 0$ and $\rho > 0$ such that the two-sided CDF increment condition in Lemma 3 holds with $F = F^A_P$ and $p = 1 - \alpha$. A sufficient condition is that F^A_P is absolutely continuous and admits a density version satisfying

$$f^A_P(x) \geq \lambda_{\min} \qquad \textit{for Lebesgue-a.e. } x \in [q_P(1-\alpha) - \rho,\, q_P(1-\alpha) + \rho].$$

If

$$\Delta^A_{\mathrm{trim}} < \lambda_{\min}\rho,$$

and, in particular, if the sufficient condition

$$1 - p_c < \lambda_{\min}\rho$$

holds, then

$$\big|q_{\mathrm{keep}}(1-\alpha) - q_P(1-\alpha)\big| \leq \frac{\Delta^A_{\mathrm{trim}}}{\lambda_{\min}} \leq \frac{1-p_c}{\lambda_{\min}}. \tag{36}$$

Proof. Apply Lemma 3 with

$$F = F^A_P, \qquad G = F^A_{P_{\mathrm{keep}}}, \qquad p = 1 - \alpha.$$

By definition of Δ^A_{trim},

$$\sup_{t \in \mathbb{R}} \big|F^A_P(t) - F^A_{P_{\mathrm{keep}}}(t)\big| = \Delta^A_{\mathrm{trim}}.$$

By Lemma 4,

$$\Delta^A_{\mathrm{trim}} \leq 1 - p_c.$$

Hence the sufficient condition $1 - p_c < \lambda_{\min}\rho$ implies

$$\Delta^A_{\mathrm{trim}} < \lambda_{\min}\rho.$$

Under this sufficient condition, or under the stated direct condition $\Delta^A_{\mathrm{trim}} < \lambda_{\min}\rho$, the hypotheses of Lemma 3 are satisfied. Consequently,

$$\big|q_{\mathrm{keep}}(1-\alpha) - q_P(1-\alpha)\big| \leq \frac{\Delta^A_{\mathrm{trim}}}{\lambda_{\min}} \leq \frac{1-p_c}{\lambda_{\min}},$$

proving (36). □

Remark 10 (Moderate trimming, aggressive trimming, and width interpretation). The condition

$$\Delta^A_{\mathrm{trim}} < \lambda_{\min}\rho \quad \text{(or the sufficient condition } 1 - p_c < \lambda_{\min}\rho)$$

is a moderate-trimming requirement. It can fail under aggressive trimming, because the retained clean mass p_c) may become too small. The local density lower bound around $q_P(1-\alpha)$ then cannot control the quantile perturbation. In that regime, Corollary 2 should be read as a diagnostic for mild-enough trimming and oracle-quantile stability, not as a blanket statement over all thresholds.

For the common symmetric interval rule

$$C_\alpha(x) = [\hat{f}(x) - \tau,\, \hat{f}(x) + \tau],$$

the population-oracle width is 2τ. Therefore, (36) immediately implies that, whenever the assumptions of the corollary hold, the oracle width perturbation induced by trimming is controlled by

$$2\big|q_{\mathrm{keep}}(1-\alpha) - q_P(1-\alpha)\big| \leq \frac{2\Delta^A_{\mathrm{trim}}}{\lambda_{\min}}.$$

F Proofs of main generic results

F.1 Proof of Proposition 1

Proof. Work conditionally on $\mathcal{G}$. Then A, S, and $t^\star$ are fixed, and the contaminated calibration observations are i.i.d. from $P^\star$. Let

$$K_i = \mathbf{1}\{S(Z_i) \le t^\star\}, \qquad \mathcal{I}_{\mathrm{keep}} = \{i : K_i = 1\}.$$

Each K_i has success probability μ_{keep}, so $N_{\mathrm{keep}} \sim \mathrm{Binomial}(m, \mu_{\mathrm{keep}})$. For any deterministic index set $I = \{i_1 < \cdots < i_n\}$ with positive probability, the event $\mathcal{I}_{\mathrm{keep}} = I$ factors coordinate wise into retained and discarded events. Independence gives

$$\mathcal{L}\big((Z_{i_1}, \ldots, Z_{i_n}) \mid \mathcal{I}_{\mathrm{keep}} = I, \mathcal{G}\big) = R^{\otimes n}.$$

Since the right-hand side does not depend on I, the same product-law statement holds after conditioning only on $N_{\mathrm{keep}} = n$. Thus, the retained observations are i.i.d. from R. Their scores are exchangeable with an independent retained-law test score $A(W_0)$, where $W_0 \sim R$.

If $n = 0$, or if $\lceil (n+1)(1-\alpha) \rceil = n+1$, the procedure sets $\hat{\tau}_\alpha = +\infty$, and retained-law coverage is one. Otherwise let

$$r_n = \lceil (n+1)(1-\alpha) \rceil \le n.$$

Attach auxiliary tie-breakers $U_0, U_1, \ldots, U_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1)$, independent of the scores. Order the pairs $(A(W_i), U_i)$ lexicographically. Let $\widetilde{R}_0$ be the rank of the test pair among the $n+1$ pairs. By exchangeability and the continuous tie-breakers,

$$\mathbb{P}\{\widetilde{R}_0 \le r_n \mid N_{\mathrm{keep}} = n, \mathcal{G}\} = \frac{r_n}{n+1} \ge 1 - \alpha.$$

Let $\hat{\tau}_\alpha$ be the ordinary inclusive r_n-th order statistic of the n retained calibration scores. We claim

$$\{\widetilde{R}_0 \le r_n\} \subseteq \{A(W_0) \le \hat{\tau}_\alpha\}.$$

Indeed, if $A(W_0) > \hat{\tau}_\alpha$, then at least r_n retained calibration scores are no larger than $\hat{\tau}_\alpha$, and hence are strictly smaller than $A(W_0)$. Their lexicographic pairs all precede the test pair, so $\widetilde{R}_0 \ge r_n + 1$, a contradiction. Therefore

$$\mathbb{P}_R\{A(W_0) \le \hat{\tau}_\alpha \mid N_{\mathrm{keep}} = n, \mathcal{G}\} \ge 1 - \alpha.$$

Averaging over N_{keep} proves

$$\mathbb{E}\{F_R^A(\hat{\tau}_\alpha) \mid \mathcal{G}\} = \mathbb{P}_R\{A(W_0) \le \hat{\tau}_\alpha \mid \mathcal{G}\} \ge 1 - \alpha.$$

For the transfer statement, set

$$\Delta_{R,P_0,+} := \sup_t \big(F_R^A(t) - F_{P_0}^A(t)\big)_+.$$

Then $F_{P_0}^A(t) \ge F_R^A(t) - \Delta_{R,P_0,+}$ for all finite t, and also at $+\infty$. Hence

$$\mathbb{P}_{P_0}\{A(Z_{\mathrm{new}}) \le \hat{\tau}_\alpha \mid \mathcal{G}\} = \mathbb{E}\{F_{P_0}^A(\hat{\tau}_\alpha) \mid \mathcal{G}\} \ge \mathbb{E}\{F_R^A(\hat{\tau}_\alpha) \mid \mathcal{G}\} - \Delta_{R,P_0,+} \ge 1 - \alpha - \Delta_{R,P_0,+}.$$

□

F.2 Proof of Theorem 1

Proof of Theorem 1. Work conditionally on $\mathcal{G}$, and suppress this conditioning in the notation. Then A, S, and $t^\star$ are fixed. The calibration observations are i.i.d. from $P^\star$, and the target test point is independent of the calibration sample. The retained law

$$R = P^\star(\,\cdot \mid S(Z) \le t^\star\,)$$

is well-defined because $\mu_{\mathrm{keep}} > 0$. We prove the identity for an arbitrary independent target law P_0, and recover the stated case by taking $P_0 = P$.

By the thinning argument in the proof of Proposition 1, conditional on $N_{\mathrm{keep}} = n$, the retained scores have the same joint law as

$$A(W_1), \ldots, A(W_n), \qquad W_1, \ldots, W_n \overset{\mathrm{iid}}{\sim} R,$$

Moreover, $N_{\mathrm{keep}} \sim \mathrm{Binomial}(m, \mu_{\mathrm{keep}})$.

Fix n. If $n = 0$ or $r_n = n + 1$, the algorithm sets $\hat{\tau}_\alpha = +\infty$, and

$$\mathbb{P}_{P_0}\{A(Z_{\mathrm{new}}) \le \hat{\tau}_\alpha \mid N_{\mathrm{keep}} = n\} = 1.$$

This is the first branch of $\Psi_n(P_0, R)$.

Suppose $1 \le r_n \le n$. Let $U_1, \ldots, U_n \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(0,1)$ and set $X_i = Q_R^A(U_i)$. By the generalised inverse transform, $X_i \sim F_R^A$. Since Q_R^A is non decreasing,

$$X_{(r_n)} \;=\; Q_R^A(U_{(r_n)})$$

almost surely, even when F_R^A has atoms. The order statistic $U_{(r_n)} \sim \mathrm{Beta}(r_n, n + 1 - r_n)$, so the inclusive retained cutoff conditional on $N_{\mathrm{keep}} = n$ has the same distribution as $Q_R^A(B_{r_n, n+1-r_n})$. Conditioning on the cutoff and using independence of the test point,

$$\mathbb{P}_{P_0}\{A(Z_{\mathrm{new}}) \le \hat{\tau}_\alpha \mid \hat{\tau}_\alpha, N_{\mathrm{keep}} = n\} = F_{P_0}^A(\hat{\tau}_\alpha),$$

hence

$$\mathbb{P}_{P_0}\{A(Z_{\mathrm{new}}) \le \hat{\tau}_\alpha \mid N_{\mathrm{keep}} = n\} = \mathbb{E}\big[F_{P_0}^A\{Q_R^A(B_{r_n, n+1-r_n})\}\big].$$

Combining the two branches and averaging over $N_{\mathrm{keep}} \sim \mathrm{Binomial}(m, \mu_{\mathrm{keep}})$ yields the target-law identity. Taking $P_0 = P$ gives (4). □

Remark 11 (Mixture upper bound on the retained-to-clean discrepancy). The componentwise upper bound used in Section 3.2 motivates the clean-side and dirty-side decomposition. It follows from the retained-law identity (3) by a short triangle-inequality argument, recorded here for reference.

Assume $p_c > 0$. Bayes' rule gives the retained-law decomposition $R = (1 - \tilde{\varepsilon}_\star)P_{\mathrm{keep}} + \tilde{\varepsilon}_\star Q_{\mathrm{keep}}$, with Q_{keep} arbitrary if $p_d = 0$. Therefore

$$F_R^A(t) - F_P^A(t) = (1 - \tilde{\varepsilon}_\star)\big(F_{P_{\mathrm{keep}}}^A(t) - F_P^A(t)\big) + \tilde{\varepsilon}_\star\big(F_{Q_{\mathrm{keep}}}^A(t) - F_P^A(t)\big),$$

and applying $(\cdot)_+$ pointwise followed by the supremum yields

$$d_{R \to P,+}^A \le (1 - \tilde{\varepsilon}_\star)\Delta_{\mathrm{trim},+}^A + \tilde{\varepsilon}_\star\, D_{Q,+}. \tag{37}$$

The two-sided variant follows analogously, with Δ_{trim}^A and D_Q replacing their one-sided counterparts. Combining (37) with Proposition 2 gives the population diagnostic lower bound stated in Section 3.2. This bound is also used in the operational certificate of Proposition 11.

F.3 Proof of Proposition 2

Proof. Work conditionally on $\mathcal{G}$, and write $d = d_{R \to P,+}^A$. By definition of d, for every score threshold t,

$$F_P^A(t) \ge F_R^A(t) - d.$$

Since $F_R^A(Q_R^A(u)) \ge u$ for the lower generalised quantile, this implies, for every $u \in (0,1)$,

$$F_P^A\{Q_R^A(u)\} \ge u - d.$$

Because a CDF is non negative, $F_P^A\{Q_R^A(u)\} \ge 0$ as well. Hence

$$F_P^A\{Q_R^A(u)\} \ge (u - d)_+.$$

Applying this pointwise inequality to the exact identity (4) gives the first inequality in (6). On the branch $r_n = n + 1$, the cutoff is $+\infty$, and both the exact conditional coverage contribution and $\psi_n(d)$ equal one. On the branch $r_n \le n$, the exact contribution is

$$\mathbb{E}\big[F_P^A\{Q_R^A(B_{r_n, n+1-r_n})\}\big] \ge \mathbb{E}\big[(B_{r_n, n+1-r_n} - d)_+\big].$$

Averaging over $N_{\mathrm{keep}} \sim \mathrm{Binomial}(m, \mu_{\mathrm{keep}})$ proves the first bound.

The closed form for $\psi_n(d)$ follows from

$$\mathbb{E}[(B-d)_+] = \mathbb{E}[B\mathbf{1}\{B > d\}] - d\,\mathbb{P}(B > d),$$

with $B \sim \mathrm{Beta}(a, b)$, together with

$$\mathbb{P}(B > d) = 1 - I_d(a, b), \qquad \mathbb{E}[B\mathbf{1}\{B > d\}] = \frac{a}{a+b}\{1 - I_d(a+1, b)\}.$$

Taking $a = r_n$ and $b = n + 1 - r_n$ gives the displayed expression.

For the second bound, if $r_n = n + 1$, then $\psi_n(d) = 1 \geq [1 - \alpha - d]_+$. If $r_n \leq n$, then

$$\psi_n(d) = \mathbb{E}[(B_{r_n, n+1-r_n} - d)_+] \geq \mathbb{E}[B_{r_n, n+1-r_n}] - d = \frac{r_n}{n+1} - d \geq 1 - \alpha - d.$$

Since $\psi_n(d) \geq 0$, this gives $\psi_n(d) \geq [1 - \alpha - d]_+$ for every n. The binomial weights sum to one, so $L_{\mathrm{fs}}(d) \geq [1 - \alpha - d]_+$. □

F.4 Proof of Theorem 2

Proof. Let $p = 1 - \alpha$. We prove the minimax identity in two directions.

Take any $(R^A, P^A) \in \mathfrak{C}_d$ and put

$$d_0 := \sup_t \big(F_R^A(t) - F_P^A(t)\big)_+ \leq d.$$

The proof of Proposition 2 is a score-law argument and applies directly to the abstract experiment defining $\mathcal{C}_{m,\mu}$. It gives

$$\mathcal{C}_{m,\mu}(R^A, P^A) \geq L_{\mathrm{fs}}^{m,\mu}(d_0).$$

For every non degenerate retained count, $\psi_n(x) = \mathbb{E}[(B_{r_n, n+1-r_n} - x)_+]$ is non increasing in x; in the degenerate branch it is identically one. Hence $L_{\mathrm{fs}}^{m,\mu}(x)$ is non increasing, and therefore

$$\mathcal{C}_{m,\mu}(R^A, P^A) \geq L_{\mathrm{fs}}^{m,\mu}(d_0) \geq L_{\mathrm{fs}}^{m,\mu}(d).$$

Taking the infimum over $\mathfrak{C}_d$ proves the "$\geq$" direction of (7).

Take R_d^A to be uniform on $[0, 1]$, so that

$$F_{R_d}^A(t) = \begin{cases} 0, & t < 0, \\ t, & 0 \leq t < 1, \\ 1, & t \geq 1. \end{cases}$$

Define the continuous target score law

$$F_{P_d}^A(t) = \begin{cases} 0, & t < d, \\ t - d, & d \leq t < 1, \\ 1 - d + d(t-1), & 1 \leq t < 2, \\ 1, & t \geq 2. \end{cases}$$

This is a valid continuous CDF: it has density one on $[d, 1]$, density d on $[1, 2]$, and total mass $(1 - d) + d = 1$. For $t \in [d, 1]$,

$$F_{R_d}^A(t) - F_{P_d}^A(t) = d.$$

For $t \in [0, d)$, the difference is $t \leq d$; for $t \in [1, 2]$, the difference is $d(2 - t) \leq d$; and outside these intervals it is non positive or zero. Hence

$$\sup_t \{F_{R_d}^A(t) - F_{P_d}^A(t)\}_+ = d,$$

so $(R_d^A, P_d^A) \in \mathfrak{C}_d$ with equality in the discrepancy constraint.

Now condition on $N = n$. If $r_n = n + 1$, the cutoff is $+\infty$ and the coverage contribution is one, equal to $\psi_n(d)$. If $r_n \leq n$, the cutoff is the r_n-th order statistic of n i.i.d. uniform variables, so

$$\widehat{q}_n \sim \mathrm{Beta}(r_n, n + 1 - r_n).$$

Since $\widehat{q}_n \in [0,1]$ almost surely,

$$F^A_{P_d}(\widehat{q}_n) = (\widehat{q}_n - d)_+,$$

and the conditional target coverage is

$$\mathbb{E}[(B_{r_n,n+1-r_n} - d)_+] = \psi_n(d).$$

Averaging these exact conditional contributions over $N \sim \mathrm{Binomial}(m,\mu)$ yields

$$\mathcal{C}_{m,\mu}(R^A_d, P^A_d) = L^{m,\mu}_{\mathrm{fs}}(d).$$

This proves the "$\leq$" direction and shows that the infimum is attained.

Because $\mu \in (0,1]$ is fixed, $N \to \infty$ in probability as $m \to \infty$. Conditional on $N = n \to \infty$, the beta order statistic $B_{r_n,n+1-r_n}$ converges in probability to p, since $r_n/(n+1) \to p$. The variables $(B_{r_n,n+1-r_n} - d)_+$ are bounded by one. Bounded convergence therefore gives convergence of the binomial mixture expectation to $(p-d)_+$. The fallback probability tends to zero, and hence

$$L^{m,\mu}_{\mathrm{fs}}(d) \to [p-d]_+ = [(1-\alpha)-d]_+.$$

□

F.5 Same-sample threshold selection

The exact retained-law identity in Theorem 1 requires the threshold $t^\star$ to be fixed before the final contaminated calibration sample is used for the conformal quantile. If the threshold is selected from the same calibration data, the retained sample no longer has the fixed-threshold product law after conditioning on the selected threshold. The rank argument in Proposition 1 therefore no longer applies.

The following result gives a finite-grid, selection-aware alternative. It is a high-probability empirical-CDF transfer bound that holds uniformly over all candidate thresholds. It therefore allows arbitrary data-dependent selection rules, at the cost of a $\sqrt{\log K/N_t}$ grid-search penalty. The result is only a certificate template: the population losses d_t must still be bounded independently to obtain a numerical guarantee.

Proposition 10 (Finite-grid same-sample threshold selection)**.** *Condition on $\mathcal{G}$. Let $\mathcal{T} = \{t_1,\dots,t_K\} \subset \overline{\mathbb{R}}$ be a finite set of candidate thresholds such that $\mu_t := P^\star(S \leq t) > 0$ for every $t \in \mathcal{T}$. For each $t \in \mathcal{T}$, define*

$$I_t = \{i : S(Z_i) \leq t\}, \qquad N_t = |I_t|, \qquad R_t = P^\star(\,\cdot \mid S \leq t\,),$$

and let $\hat{\tau}_t$ be the trimmed conformal cutoff computed from I_t by the inclusive quantile rule of Algorithm 1. Define the per-threshold retained-to-clean loss

$$d_t = d^A_{R_t \to P,+} = \sup_a \big(F^A_{R_t}(a) - F^A_P(a)\big)_+.$$

Fix $\beta \in (0,1)$. Set $r_t = \lceil (N_t+1)(1-\alpha) \rceil$ for all t, and, when $N_t \geq 1$, set

$$\eta_t = \sqrt{\frac{\log(2K/\beta)}{2N_t}}.$$

Let $\widehat{t}$ be any data-dependent rule taking values in $\mathcal{T}$. It may be chosen after inspecting all calibration scores and all candidate cutoffs.

Then, with conditional probability at least $1-\beta$ over the final contaminated calibration sample, the following holds simultaneously for every $t \in \mathcal{T}$ with $N_t \geq 1$ and $r_t \leq N_t$,

$$\mathbb{P}_P\big\{A(Z_{\mathrm{new}}) \leq \hat{\tau}_t \big| Z_1,\dots,Z_m,\mathcal{G}\big\} \geq \frac{r_t}{N_t} - d_t - \eta_t. \tag{38}$$

For thresholds with $N_t = 0$ or $r_t = N_t + 1$, the algorithm sets $\hat{\tau}_t = +\infty$, and the conditional clean coverage equals one. On the same event, the bound (38) holds at $t = \widehat{t}$ whenever the selected threshold is non degenerate. If it is degenerate, the conditional clean coverage is one. No independence between $\widehat{t}$ and the calibration sample is required, because the event holds uniformly over $\mathcal{T}$.

The loss d_t is a population retained-to-clean discrepancy for the candidate threshold. It is not directly observable. Proposition 10 is therefore selection-aware for the empirical retained CDF, but it is not by itself a data-only coverage certificate. To obtain a numerical guarantee, the losses $\{d_t\}_{t\in\mathcal{T}}$ must be bounded using independent information. This can be done, for example, with a clean audit sample (Proposition 12) or with simultaneous independent CDF bounds over the grid.

Proof of Proposition 10. Work conditionally on $\mathcal{G}$ throughout. Then the candidate grid is fixed, and $Z_1, \dots, Z_m$ are i.i.d. from $P^\star$.

Fix $t \in \mathcal{T}$. Since $\mu_t > 0$, the retained law

$$R_t = P^\star(\cdot \mid S \le t)$$

is well-defined. Conditional on $N_t = n \ge 1$, the retained observations $\{Z_i : i \in I_t\}$ are i.i.d. from R_t by independent thinning. Hence their retained-score empirical CDF

$$\widehat{F}_t^A(a) := \frac{1}{N_t} \sum_{i\in I_t} \mathbf{1}\{A(Z_i) \le a\}, \qquad N_t \ge 1,$$

satisfies, for every $n \ge 1$,

$$\begin{aligned}\mathbb{P}\left(\sup_a \left|\widehat{F}_t^A(a) - F_{R_t}^A(a)\right| > \sqrt{\frac{\log(2K/\beta)}{2n}} \;\middle|\; N_t = n, \mathcal{G} \right) \\ \le 2\exp\left\{-2n \cdot \frac{\log(2K/\beta)}{2n}\right\} = \frac{\beta}{K},\end{aligned}$$

by the two-sided Dvoretzky–Kiefer–Wolfowitz inequality with Massart's constant. Define the per-threshold good event

$$\mathcal{E}_t := \{N_t = 0\} \cup \left\{ N_t \ge 1, \sup_a \left|\widehat{F}_t^A(a) - F_{R_t}^A(a)\right| \le \sqrt{\frac{\log(2K/\beta)}{2N_t}} \right\}.$$

Equivalently, $\mathcal{E}_t^c$ is the event where $N_t \ge 1$ and the empirical-CDF deviation exceeds the random threshold η_t. Averaging the conditional DKW bound over the possible positive retained counts gives

$$\begin{aligned}\mathbb{P}(\mathcal{E}_t^c \mid \mathcal{G}) &= \sum_{n=1}^m \mathbb{P}(N_t = n \mid \mathcal{G}) \mathbb{P}\left(\sup_a \left|\widehat{F}_t^A(a) - F_{R_t}^A(a)\right| > \sqrt{\frac{\log(2K/\beta)}{2n}} \;\middle|\; N_t = n, \mathcal{G} \right) \\ &\le \frac{\beta}{K}\, \mathbb{P}(N_t \ge 1 \mid \mathcal{G}) \le \frac{\beta}{K}.\end{aligned}$$

This is where the zero-retained-count case must be included in the good event. Otherwise, a threshold with $N_t = 0$ would incorrectly make $\mathcal{E}_t$ fail.

Let

$$\mathcal{E} := \bigcap_{t\in\mathcal{T}} \mathcal{E}_t.$$

By the preceding display and the union bound,

$$\mathbb{P}(\mathcal{E}^c \mid \mathcal{G}) \le \sum_{t\in\mathcal{T}} \mathbb{P}(\mathcal{E}_t^c \mid \mathcal{G}) \le K \cdot \frac{\beta}{K} = \beta.$$

Thus $\mathcal{E}$ holds with conditional probability at least $1 - \beta$. On $\mathcal{E}$, for every candidate threshold with $N_t \ge 1$,

$$\sup_a \left|\widehat{F}_t^A(a) - F_{R_t}^A(a)\right| \le \eta_t. \tag{39}$$

Work on $\mathcal{E}$, and fix any $t \in \mathcal{T}$ with $N_t \ge 1$ and $r_t \le N_t$. The cutoff $\hat{\tau}_t$ is the r_t-th order statistic of the N_t retained scores. Hence the inclusive empirical CDF satisfies

$$\widehat{F}_t^A(\hat{\tau}_t) \ge \frac{r_t}{N_t}.$$

Combining this with (39),

$$F^A_{R_t}(\hat{\tau}_t) \geq \widehat{F}^A_t(\hat{\tau}_t) - \eta_t \geq \frac{r_t}{N_t} - \eta_t.$$

By definition of d_t, for every finite score threshold a,

$$F^A_P(a) \geq F^A_{R_t}(a) - d_t.$$

Therefore

$$F^A_P(\hat{\tau}_t) \geq F^A_{R_t}(\hat{\tau}_t) - d_t \geq \frac{r_t}{N_t} - d_t - \eta_t.$$

Given $\mathcal{G}$, the clean test point is independent of the calibration sample. The cutoff $\hat{\tau}_t$ is measurable with respect to $Z_1, \ldots, Z_m$ and $\mathcal{G}$. Hence

$$\mathbb{P}_P\big\{A(Z_{\text{new}}) \leq \hat{\tau}_t \,\big|\, Z_1, \ldots, Z_m, \mathcal{G}\big\} = F^A_P(\hat{\tau}_t),$$

which proves (38).

If $N_t = 0$ or $r_t = N_t + 1$, the algorithm sets $\hat{\tau}_t = +\infty$, and the conditional clean coverage is $F^A_P(+\infty) = 1$. The event $\mathcal{E}$ is defined before selecting a threshold and holds uniformly over the finite grid. Hence, after $\mathcal{E}$ occurs, any measurable rule $\widehat{t} \in \mathcal{T}$ inherits the corresponding bound at its selected threshold when the selected branch is non degenerate. If it is degenerate, it inherits the trivial coverage-one statement. This holds regardless of how $\widehat{t}$ depends on the calibration sample. □

Corollary 3 (Finite-library score-threshold selection)**.** *Condition on $\mathcal{G}$. Let*

$$\mathcal{H} = \{(S_h, t_h) : h = 1, \ldots, H\}$$

be a finite library of measurable score-threshold pairs, fixed conditionally on $\mathcal{G}$. Assume that the nonconformity score A is also fixed conditionally on $\mathcal{G}$, and that

$$\mu_h = P^\star(S_h \leq t_h) > 0$$

for every h. For each candidate, let N_h be the retained count, $\widehat{\tau}_h$ the inclusive conformal cutoff, and define

$$R_h = P^\star(\cdot \mid S_h \leq t_h), \qquad d_h = d^A_{R_h \to P,+}.$$

Fix $\beta \in (0,1)$. For each h with $N_h \geq 1$, set

$$r_h = \lceil (N_h + 1)(1 - \alpha) \rceil, \qquad \eta_h = \sqrt{\frac{\log(2H/\beta)}{2N_h}}.$$

Then, with conditional probability at least $1 - \beta$ over the final calibration sample, simultaneously for all candidates satisfying $N_h \geq 1$ and $r_h \leq N_h$,

$$\mathbb{P}_P\left(A(Z_{\text{new}}) \leq \widehat{\tau}_h \mid Z_{1:m}, \mathcal{G}\right) \geq \frac{r_h}{N_h} - d_h - \eta_h.$$

Consequently, the same inequality holds for any data-dependent selection $\widehat{h} \in \{1, \ldots, H\}$ whose selected cutoff is finite. For a selected candidate with cutoff $+\infty$, the conditional coverage equals one.

Proof. Fix a candidate h. Conditional on its retained count $N_h = n \geq 1$, the retained observations are i.i.d. from R_h.

The Dvoretzky–Kiefer–Wolfowitz inequality gives

$$\mathbb{P}\left(\sup_a \left|\widehat{F}^A_h(a) - F^A_{R_h}(a)\right| > \eta_h \mid N_h = n, \mathcal{G}\right) \leq \frac{\beta}{H}.$$

Averaging over the retained count and applying a union bound over the H candidates gives a simultaneous event with probability at least $1 - \beta$.

On this event, every finite selected cutoff obeys

$$F^A_{R_h}(\widehat{\tau}_h) \geq \widehat{F}^A_h(\widehat{\tau}_h) - \eta_h \geq \frac{r_h}{N_h} - \eta_h.$$

By the definition of d_h,

$$F_P^A(\widehat{\tau}_h) \geq F_{R_h}^A(\widehat{\tau}_h) - d_h.$$

Combining these inequalities proves the result. The event holds for the entire library before selection, so it also holds for any data-dependent choice from the library. □

F.6 Proof of Proposition 3

Proof of Proposition 3. The proof has three parts: the covariance identity itself, the homoscedastic specialisation, and the approximate residual bound under model misspecification.

For any $t \in \mathbb{R}$,

$$F_{P_{\text{keep}}}^A(t) = P\big(A(Z) \leq t \big| K = 1\big) = \frac{\mathbb{E}_P\big[\mathbf{1}\{A(Z) \leq t\}K\big]}{p_c},$$

hence

$$\begin{aligned} F_{P_{\text{keep}}}^A(t) - F_P^A(t) &= \frac{\mathbb{E}_P[\mathbf{1}\{A(Z) \leq t\}\, K]}{p_c} - \mathbb{E}_P[\mathbf{1}\{A(Z) \leq t\}] \\ &= \frac{\mathrm{Cov}_P\big(\mathbf{1}\{A(Z) \leq t\},\, K\big)}{p_c}, \end{aligned}$$

which establishes the pointwise covariance identity. Taking positive parts and the supremum over t gives the one-sided form $\Delta_{\text{trim},+}^A = p_c^{-1} \sup_t \{\mathrm{Cov}_P(\mathbf{1}\{A(Z) \leq t\}, K)\}_+$, and taking absolute values gives the two-sided form $\Delta_{\text{trim}}^A = p_c^{-1} \sup_t |\mathrm{Cov}_P(\mathbf{1}\{A(Z) \leq t\}, K)|$. If $A(Z)$ and K are independent under P, every covariance in these displays vanishes, so $F_{P_{\text{keep}}}^A \equiv F_P^A$ and $\Delta_{\text{trim}}^A = \Delta_{\text{trim},+}^A = \Delta_{\text{trim},-}^A = 0$.

Suppose $Y = f^\star(X) + \sigma\xi$ with $\xi \perp\!\!\!\perp X$, $\hat{f} = f^\star$, $A(X,Y) = |Y - \hat{f}(X)|$, and S covariate-only. Then

$$A(X,Y) = |Y - \hat{f}(X)| = \sigma|\xi|,$$

which depends on X only through ξ. Since $\xi \perp\!\!\!\perp X$, the variable $A(X,Y)$ is independent of any covariate-only retention indicator $K = \mathbf{1}\{S_X(X) \leq t^\star\}$ under P. Then gives $\Delta_{\text{trim}}^A = 0$.

Now suppose only that $Y = f^\star(X) + \sigma\xi$ with $\xi \perp\!\!\!\perp X$, but allow $\hat{f} \neq f^\star$ with $\|f^\star - \hat{f}\|_\infty \leq \eta$. Define $b(X) := f^\star(X) - \hat{f}(X)$ and $A_0 := \sigma|\xi|$, so that $\|b\|_\infty \leq \eta$ and

$$A(X,Y) \;=\; |\sigma\xi + b(X)|.$$

The reverse triangle inequality gives $|A - A_0| \leq |b(X)| \leq \eta$. Under the same covariate-only retention condition, $A_0 \perp\!\!\!\perp K$ under P, so the law of A_0 is unaffected by conditioning on $\{K = 1\}$:

$$F_{P_{\text{keep}}}^{A_0} \;\equiv\; F_P^{A_0}.$$

For any t, the inequality $A \leq t$ implies $A_0 \leq t + \eta$, and conversely $A_0 \leq t - \eta$ implies $A \leq t$. Hence

$$F_{P_{\text{keep}}}^A(t) \;\leq\; F_{P_{\text{keep}}}^{A_0}(t+\eta) \;=\; F_P^{A_0}(t+\eta), \qquad F_P^A(t) \;\geq\; F_P^{A_0}(t-\eta),$$

which give

$$F_{P_{\text{keep}}}^A(t) - F_P^A(t) \;\leq\; F_P^{A_0}(t+\eta) - F_P^{A_0}(t-\eta) \;=\; \mathbb{P}\{A_0 \in [t-\eta, t+\eta]\}.$$

Reversing the roles of P_{keep} and P in the same chain of inequalities gives the symmetric upper bound on $F_P^A(t) - F_{P_{\text{keep}}}^A(t)$. Taking the supremum over t yields the residual smoothness bound

$$\Delta_{\text{trim}}^A \leq \omega_{A_0}(\eta) := \sup_t \mathbb{P}A_0 \in [t-\eta, t+\eta].$$

If A_0 has density bounded by M, every interval of length 2η has probability at most $2M\eta$. Hence $\omega_{A_0}(\eta) \leq 2M\eta$. □

F.7 Proof of Proposition 4

Proof of Proposition 4. Work conditionally on $\mathcal{G}$. Since $p_c > 0$ and $\varepsilon \in [0, 1)$, the denominator in

$$\tilde{\varepsilon}_\star = \frac{\varepsilon p_d}{(1-\varepsilon)p_c + \varepsilon p_d}$$

is positive. For fixed $p_c > 0$ and $\varepsilon \in [0, 1)$, define

$$g(x) = \frac{\varepsilon x}{(1-\varepsilon)p_c + \varepsilon x}, \qquad x \geq 0.$$

Then

$$g'(x) = \frac{\varepsilon(1-\varepsilon)p_c}{\{(1-\varepsilon)p_c + \varepsilon x\}^2} \geq 0,$$

so g is nondecreasing in the dirty retention probability. If $p_d \leq \lambda p_c$, then

$$\tilde{\varepsilon}_\star = g(p_d) \leq g(\lambda p_c) = \frac{\varepsilon\lambda p_c}{(1-\varepsilon)p_c + \varepsilon\lambda p_c} = \frac{\varepsilon\lambda}{1-\varepsilon+\varepsilon\lambda}.$$

In particular, if $p_d/p_c \to 0$, take $\lambda = p_d/p_c$ to obtain $\tilde{\varepsilon}_\star \to 0$. If $p_d = p_c$, then

$$\tilde{\varepsilon}_\star = \frac{\varepsilon p_c}{(1-\varepsilon)p_c + \varepsilon p_c} = \varepsilon.$$

This proves the proposition. □

F.8 Proofs for independent adaptive threshold selection

Proof. Work conditionally on $\mathcal{G} \vee \mathcal{H}$. Under the theorem's assumptions, A, S, and $\widehat{t}$ are fixed after this conditioning. The contaminated calibration observations remain i.i.d. from $P^\star$ and independent of the clean test point.

Therefore, Theorem 1 applies verbatim with $t^\star = \widehat{t}$. It gives the untruncated bound $1 - \alpha - d^A_{R_{\widehat{t}} \to P,+}$. Non-negativity of coverage gives the positive-part version stated in the theorem, on the event $\mu_{\mathrm{keep}}(\widehat{t}) > 0$.

The componentwise mixture upper-bound diagnostic additionally requires $p_c(\widehat{t}) > 0$. The exact identity also follows, using generalised retained-score quantiles and without any continuity requirement. □

Proof. Work on $\mathcal{E}_{\mathrm{tune}} \cap \{\widehat{\mathcal{T}}_\eta \neq \varnothing\}$. Since $\widehat{t} \in \widehat{\mathcal{T}}_\eta$, $\widehat{B}(\widehat{t}) \leq \eta$. The uniform tuning error gives

$$B(\widehat{t}) \leq \widehat{B}(\widehat{t}) + r_B \leq \eta + r_B,$$

proving (10). Since $d^A_{R_{\widehat{t}} \to P,+} \leq B(\widehat{t})$, Theorem 3 yields the stated conditional coverage bound on this event.

For efficiency, let $t \in \mathcal{T}$ satisfy $B(t) \leq \eta - r_B$. On $\mathcal{E}_{\mathrm{tune}}$, $\widehat{B}(t) \leq B(t) + r_B \leq \eta$, so $t \in \widehat{\mathcal{T}}_\eta$. Therefore $\widehat{W}(\widehat{t}) \leq \widehat{W}(t)$. Applying the uniform efficiency error twice gives

$$W(\widehat{t}) \leq \widehat{W}(\widehat{t}) + r_W \leq \widehat{W}(t) + r_W \leq W(t) + 2r_W.$$

Taking the infimum over t with $B(t) \leq \eta - r_B$ proves (11).

For the marginal statement, on $\mathcal{E}_{\mathrm{tune}} \cap \{\widehat{\mathcal{T}}_\eta \neq \varnothing\}$ the conditional coverage is at least $[1-\alpha-\eta-r_B]_+$. The complement of this event has probability at most $\beta + \beta_\varnothing$, and coverage probabilities are nonnegative. Therefore, the marginal coverage is at least

$$[1-\beta-\beta_\varnothing]_+[1-\alpha-\eta-r_B]_+ \geq 1-\alpha-\eta-r_B-\beta-\beta_\varnothing.$$

If $\widehat{\mathcal{T}}_\eta$ is empty, the proposition explicitly falls back to t_{fb}. Theorem 3 still gives the fixed-threshold guarantee for this fallback. A certified fallback loss can then replace the crude probability subtraction. □

F.9 Componentwise certificate interface

The diagnostic bounds in Section 3.2 are population quantities for the realised auxiliary stage. They become numerical certificates only when their ingredients are bounded by independent information. The following proposition formalises the componentwise route. Given independent bounds on the retention probabilities, the clean trimming distortion, and the retained dirty discrepancy, it converts the mixture upper-bound diagnostic into a deterministic clean-coverage lower bound. The result is stated in one-sided form to match the scalar clean-transfer theorem. It also allows the contamination fraction to be known only through an upper bound.

Proposition 11 (One-sided certificate interface)**.** *Under the setup of Section 3.1, suppose that* $\varepsilon \le \varepsilon_{\max} < 1$ *and that deterministic or* $\mathcal{G}$*-measurable quantities*

$$L_c, \quad U_d, \quad B_{\Delta,+}, \quad B_{Q,+}$$

satisfy

$$0 < L_c \le p_c, \qquad 0 \le p_d \le U_d, \qquad \Delta^A_{\mathrm{trim},+} \le B_{\Delta,+}, \qquad D_{Q,+} \le B_{Q,+} \le 1.$$

Define

$$\bar{\varepsilon} \;:=\; \frac{\varepsilon_{\max} U_d}{(1-\varepsilon_{\max})L_c + \varepsilon_{\max} U_d}.$$

Then, conditionally on $\mathcal{G}$*,*

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) \;\ge\; \big[\,1 - \alpha - B_{\Delta,+} - \bar{\varepsilon}\max\{B_{Q,+} - B_{\Delta,+},\, 0\}\,\big]_+. \tag{40}$$

A simpler but looser consequence is

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) \;\ge\; \big[\,1 - \alpha - B_{\Delta,+} - \bar{\varepsilon}\, B_{Q,+}\,\big]_+. \tag{41}$$

If the certificates are non-random, the conditioning can be dropped by averaging. If they are $\mathcal{G}$*-measurable and hold only on a high-probability auxiliary event, the marginalised version in Corollary 1 applies.*

The worst-case dirty-side choice $B_{Q,+} = 1$ is always valid, but often conservative. Smaller values of $B_{Q,+}$, or non-trivial bounds on p_d, require additional information about the contamination mechanism. The proposition is intentionally an *interface*. It states which independent inputs suffice for a numerical certificate, but does not construct those inputs from the contaminated calibration sample.

Proof of Proposition 11. Work conditionally on $\mathcal{G}$. The argument has three steps: a starting bound from retained-law transfer, a monotonicity bound on the retained mixture coefficient, and a componentwise bound on the mixture loss.

By Proposition 2 and the mixture upper bound on $d^A_{R\to P,+}$ derived in Section 3.2,

$$\mathbb{P}\big\{Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \mid \mathcal{G}\big\} \;\ge\; 1 - \alpha - \big[(1-\tilde{\varepsilon}_\star)\Delta^A_{\mathrm{trim},+} + \tilde{\varepsilon}_\star D_{Q,+}\big],$$

and non-negativity of coverage gives the corresponding positive-part bound. It remains to upper-bound the bracketed loss using the certificate inputs.

For fixed non-negative p_d and positive p_c, the map $e \mapsto ep_d/[(1-e)p_c + ep_d]$ is non-decreasing on $[0,1)$. Combined with $p_c \ge L_c$, $p_d \le U_d$, and $\varepsilon \le \varepsilon_{\max}$, this gives

$$\tilde{\varepsilon}_\star \;=\; \frac{\varepsilon p_d}{(1-\varepsilon)p_c + \varepsilon p_d} \;\le\; \frac{\varepsilon_{\max} U_d}{(1-\varepsilon_{\max})L_c + \varepsilon_{\max} U_d} \;=\; \bar{\varepsilon}.$$

Applying the distortion certificates $\Delta^A_{\mathrm{trim},+} \le B_{\Delta,+}$ and $D_{Q,+} \le B_{Q,+}$ and rearranging,

$$(1-\tilde{\varepsilon}_\star)\Delta^A_{\mathrm{trim},+} + \tilde{\varepsilon}_\star D_{Q,+} \;\le\; (1-\tilde{\varepsilon}_\star)B_{\Delta,+} + \tilde{\varepsilon}_\star B_{Q,+} \;=\; B_{\Delta,+} + \tilde{\varepsilon}_\star(B_{Q,+} - B_{\Delta,+}).$$

If $B_{Q,+} \ge B_{\Delta,+}$, monotonicity in $\tilde{\varepsilon}_\star$ and Step 2 give an upper bound of $B_{\Delta,+} + \bar{\varepsilon}(B_{Q,+} - B_{\Delta,+})$. If $B_{Q,+} < B_{\Delta,+}$, the second term is non-positive and the upper bound is simply $B_{\Delta,+}$. In either case,

$$(1-\tilde{\varepsilon}_\star)\Delta^A_{\mathrm{trim},+} + \tilde{\varepsilon}_\star D_{Q,+} \;\le\; B_{\Delta,+} + \bar{\varepsilon}\max\{B_{Q,+} - B_{\Delta,+},\, 0\},$$

which proves (40) after taking positive parts. The simpler bound (41) follows from $\max\{B_{Q,+} - B_{\Delta,+}, 0\} \le B_{Q,+}$. □

F.10 Audit-based certificates

The componentwise interface in Proposition 11 certifies the prediction set indirectly, by bounding each ingredient of the mixture decomposition. When an independent clean labelled audit sample is available, a tighter and more direct route is to certify the realised prediction set itself. The following proposition gives two such certificates. Exact binomial inversion for the final selected set, which is agnostic to how the set was constructed. And a one-sided Kolmogorov certificate for any data-dependent cutoff, expressed through an audited score-CDF gap.

Proposition 12 (Final-set and score-CDF audit certificates)**.** *Let $\mathcal{A}$ be a sigma-field containing all data and randomness used to construct a prediction set-valued map $\widehat{C}$, and suppose an audit sample $(X_j^{\mathrm{aud}}, Y_j^{\mathrm{aud}})_{j=1}^{n_{\mathrm{aud}}} \overset{\mathrm{iid}}{\sim} P$ is independent of $\mathcal{A}$. Define*

$$p_{\widehat{C}} = P\{Y \in \widehat{C}(X) \mid \mathcal{A}\}, \qquad H_j = \mathbf{1}\{Y_j^{\mathrm{aud}} \in \widehat{C}(X_j^{\mathrm{aud}})\}, \qquad M = \sum_{j=1}^{n_{\mathrm{aud}}} H_j.$$

For $\beta \in (0,1)$, set

$$L_{\mathrm{bin}}(M, n_{\mathrm{aud}}, \beta) = \begin{cases} 0, & M = 0, \\ \mathrm{BetaInv}\{\beta;\, M,\, n_{\mathrm{aud}} - M + 1\}, & M \geq 1, \end{cases}$$

where $\mathrm{BetaInv}\{\beta; a, b\}$ is the β-quantile of a $\mathrm{Beta}(a,b)$ distribution. Then, conditionally on $\mathcal{A}$,

$$\mathbb{P}\big\{p_{\widehat{C}} \geq L_{\mathrm{bin}}(M, n_{\mathrm{aud}}, \beta) \,\big|\, \mathcal{A}\big\} \;\geq\; 1 - \beta. \tag{42}$$

Furthermore, let I be any selected calibration index set of size $k \geq 1$, $\mathcal{A}$-measurable, and let

$$\widehat{F}_I^A(a) = \frac{1}{k} \sum_{i \in I} \mathbf{1}\{A(Z_i) \leq a\}.$$

Let $A_j^{\mathrm{aud}} = A(X_j^{\mathrm{aud}}, Y_j^{\mathrm{aud}})$ and $\widehat{F}_{\mathrm{aud}}^A$ be their empirical CDF, and suppose $c^+_{n_{\mathrm{aud}},\beta}$ satisfies

$$\mathbb{P}\left(\sup_a \big(\widehat{F}_{\mathrm{aud}}^A(a) - F_P^A(a) \big) > c^+_{n_{\mathrm{aud}},\beta} \;\middle|\; \mathcal{A} \right) \;\leq\; \beta.$$

Then, with conditional probability at least $1-\beta$, every $\mathcal{A}$-measurable cutoff $\hat{\tau}$ satisfies

$$P\{A(Z_{\mathrm{new}}) \leq \hat{\tau} \mid \mathcal{A}\} \;\geq\; \widehat{F}_I^A(\hat{\tau}) - \sup_a \big\{ \widehat{F}_I^A(a) - \widehat{F}_{\mathrm{aud}}^A(a) \big\}_+ - c^+_{n_{\mathrm{aud}},\beta}. \tag{43}$$

In particular, if $\hat{\tau}$ is the r-th order statistic of $\{A(Z_i) : i \in I\}$, then $\widehat{F}_I^A(\hat{\tau}) \geq r/k$, yielding the explicit bound

$$P\{A(Z_{\mathrm{new}}) \leq \hat{\tau} \mid \mathcal{A}\} \;\geq\; \frac{r}{k} - \sup_a \big\{ \widehat{F}_I^A(a) - \widehat{F}_{\mathrm{aud}}^A(a) \big\}_+ - c^+_{n_{\mathrm{aud}},\beta}.$$

The two certificates serve complementary purposes. The binomial inversion certificate (42) is exact. It applies to *any* prediction-set construction, including fixed-threshold trimming, same-sample threshold selection, Stein thinning, and other selected-subset procedures. The only requirement is that the audit sample is independent of all selection decisions.

The KS certificate (43) is asymptotic in n_{aud}, but it is more flexible. It directly certifies the realised score CDF. This certificate can then be combined with empirical-CDF gaps. When the selected calibration scores are well aligned with the audit scores, this can yield sharper guarantees.

Proof of Proposition 12. We prove the two certificates separately.

Conditional on $\mathcal{A}$, the prediction-set map $\widehat{C}$ is fixed and the audit indicators $H_1, \ldots, H_{n_{\mathrm{aud}}}$ are i.i.d. Bernoulli with success probability $p_{\widehat{C}}$. The lower endpoint $L_{\mathrm{bin}}(M, n_{\mathrm{aud}}, \beta)$ is the standard one-sided Clopper–Pearson inversion of the binomial model. For each fixed $p \in [0,1]$, the probability under $\mathrm{Binomial}(n_{\mathrm{aud}}, p)$ that the inverted lower endpoint exceeds p is at most β. Applying this with $p = p_{\widehat{C}}$ gives (42).

Work on the audit-side event

$$\mathcal{E}_{\mathrm{aud}} := \Big\{ \sup_a \big(\widehat{F}^A_{\mathrm{aud}}(a) - F^A_P(a) \big) \le c^+_{n_{\mathrm{aud}},\beta} \Big\},$$

which has conditional probability at least $1-\beta$ by assumption. On $\mathcal{E}_{\mathrm{aud}}$, for every a,

$$F^A_P(a) \;\ge\; \widehat{F}^A_{\mathrm{aud}}(a) - c^+_{n_{\mathrm{aud}},\beta},$$

so for every a,

$$\widehat{F}^A_I(a) - F^A_P(a) \;\le\; \widehat{F}^A_I(a) - \widehat{F}^A_{\mathrm{aud}}(a) + c^+_{n_{\mathrm{aud}},\beta}.$$

Taking suprema and positive parts,

$$d^A_{I\to P,+} \;:=\; \sup_a \big\{ \widehat{F}^A_I(a) - F^A_P(a) \big\}_+ \;\le\; \sup_a \big\{ \widehat{F}^A_I(a) - \widehat{F}^A_{\mathrm{aud}}(a) \big\}_+ + c^+_{n_{\mathrm{aud}},\beta}.$$

By the definition of $d^A_{I\to P,+}$, every $\mathcal{A}$-measurable cutoff $\hat\tau$ satisfies $F^A_P(\hat\tau) \ge \widehat{F}^A_I(\hat\tau) - d^A_{I\to P,+}$. Substituting yields (43). The explicit-rank bound follows from the inclusive empirical-CDF inequality $\widehat{F}^A_I(\hat\tau) \ge r/k$ when $\hat\tau$ is the r-th selected order statistic. □

F.11 Upper-bound, efficiency, and marginalisation statements

Proof. Work conditionally on $\mathcal{G}$. We suppress this conditioning in the notation. Define the retained mixture law as

$$R := P^\star(\,,\cdot \mid S(Z) \le t^\star).$$

By the same conditional-thinning argument used in Theorem 1, the following holds. Conditional on $N_{\mathrm{keep}} = n$, the retained calibration scores are i.i.d. draws from the law of $A(W)$, where $W \sim R$. The clean test score used for retained-law evaluation is also an independent draw from the same score law.

Fix n. Suppose

$$r_n := \lceil (n+1)(1-\alpha) \rceil \le n.$$

Under the no-ties assumption, we obtain the exact retained-law identity. The same identity also holds if the procedure uses randomised threshold tie-breaking in a consistent way.

If ties occur at the cutoff, the situation is different. The ordinary inclusive rule may have retained-law coverage strictly above the randomised-rank value. Therefore, the upper-bound statement requires either continuity of F^A_R, or consistent randomised tie-breaking.

$$\mathbb{P}_R(A_0 \le \hat\tau_\alpha \mid N_{\mathrm{keep}} = n) = \frac{r_n}{n+1}.$$

Since

$$r_n \le (n+1)(1-\alpha) + 1,$$

this is at most

$$1 - \alpha + \frac{1}{n+1}.$$

If $r_n = n+1$, then $\hat\tau_\alpha = +\infty$ and the retained-law coverage is one. In this case $r_n = n+1$ implies

$$(n+1)(1-\alpha) > n,$$

hence $\alpha < 1/(n+1)$, and therefore

$$1 \le 1 - \alpha + \frac{1}{n+1}.$$

Thus, for every $n \ge 0$,

$$\mathbb{E}\big[F^A_R(\hat\tau_\alpha) \mid N_{\mathrm{keep}} = n \big] \le 1 - \alpha + \frac{1}{n+1}.$$

Averaging over N_{keep} gives

$$\mathbb{E}\big[F^A_R(\hat\tau_\alpha) \big] \le 1 - \alpha + \mathbb{E}\left[\frac{1}{N_{\mathrm{keep}} + 1} \right].$$

For the randomised lexicographic rule, the same argument is applied to the pair $T = (A, U)$. Define the pair CDF

$$G_M(a, u) = M\{A < a\} + uM\{A = a\}, \qquad u \in [0, 1].$$

Then the randomised retained-rank calculation gives the same upper bound for $\mathbb{E}[G_R(\widehat{T})]$. Moreover,

$$G_P(a, u) - G_R(a, u) = (1 - u)\{F_P^A(a^-) - F_R^A(a^-)\} + u\{F_P^A(a) - F_R^A(a)\} \le d^A_{P\to R,+},$$

so the target-law comparison below is unchanged. In the continuous no-ties case, this reduces to the scalar CDF comparison.

The pointwise one-sided comparison gives, for all $t \in \mathbb{R} \cup \{+\infty\}$,

$$F_P^A(t) \le F_R^A(t) + d^A_{P\to R,+}.$$

Therefore

$$\mathbb{P}\big(Y_{\text{new}} \in C_\alpha(X_{\text{new}}) \mid \mathcal{G}\big) = \mathbb{E}\big[F_P^A(\hat{\tau}_\alpha)\big] \le 1 - \alpha + d^A_{P\to R,+} + \mathbb{E}\left[\frac{1}{N_{\text{keep}} + 1}\right].$$

Since

$$d^A_{P\to R,+} \le \delta^{\text{ref}}_{\text{mix},-} \le \delta^{\text{ref}}_{\text{mix}}(D_Q),$$

the two subsequent upper bounds in (12) follow.

Finally, $N_{\text{keep}} \sim \text{Binomial}(m, \mu_{\text{keep}})$. Since $\mu_{\text{keep}} > 0$,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N_{\text{keep}} + 1}\right] &= \sum_{n=0}^{m} \frac{1}{n+1}\binom{m}{n}\mu_{\text{keep}}^n(1 - \mu_{\text{keep}})^{m-n} \\
&= \frac{1}{(m+1)\mu_{\text{keep}}} \sum_{n=0}^{m} \binom{m+1}{n+1}\mu_{\text{keep}}^{n+1}(1 - \mu_{\text{keep}})^{m-n} \\
&= \frac{1 - (1 - \mu_{\text{keep}})^{m+1}}{(m+1)\mu_{\text{keep}}}.
\end{aligned}$$

This proves (12). □

Proof. Write $R = P^\star(\cdot \mid S(Z) \le t^\star)$, and let

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i \in \mathcal{I}_{\text{keep}}} \mathbf{1}\{A_i \le t\}$$

be the empirical CDF of the retained scores when $N_{\text{keep}} = n \ge 1$. Conditional on $N_{\text{keep}} = n$, the retained scores are i.i.d. from F_R^A. Hence the Dvoretzky–Kiefer–Wolfowitz inequality gives [Dvoretzky et al., 1956, Massart, 1990]

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} |\widehat{F}_n(t) - F_R^A(t)| > \eta_{n_0,\beta} \,\middle|\, N_{\text{keep}} = n\right) \le 2e^{-2n\eta^2_{n_0,\beta}} \le 2e^{-2n_0\eta^2_{n_0,\beta}} = \beta$$

for every $n \ge n_0$. Therefore the event

$$\mathcal{E}_{n_0,\beta} := \left\{N_{\text{keep}} \ge n_0 \text{ and } \sup_{t\in\mathbb{R}} |\widehat{F}_{N_{\text{keep}}}(t) - F_R^A(t)| \le \eta_{n_0,\beta}\right\}$$

has probability at least $1 - \mathbb{P}(N_{\text{keep}} < n_0) - \beta$.

Work on $\mathcal{E}_{n_0,\beta}$, and write $n = N_{\text{keep}}$. Since $n \ge n_0 \ge \lceil 1/\alpha \rceil - 1$, the conformal index

$$r_n := \lceil (n+1)(1 - \alpha) \rceil$$

satisfies $r_n \le n$, so $\hat{\tau}_\alpha$ is a finite order statistic. Let $p = 1 - \alpha$, $q = q_P(p)$, and

$$\Gamma = \Gamma_{n_0,\beta}.$$

By assumption, $h := \Gamma/\lambda_{\min} < \rho$. The local CDF-increment condition from Lemma 3 gives

$$F_P^A(q + h) \ge p + \Gamma, \qquad F_P^A(q - h) \le p - \Gamma.$$

The retained-to-clean Kolmogorov comparison gives

$$\sup_{t\in\mathbb{R}} |F_R^A(t) - F_P^A(t)| \le \delta_{\mathrm{mix}}^{\mathrm{ref}}(D_Q).$$

Hence

$$F_R^A(q+h) \ge p + \eta_{n_0,\beta} + \frac{2}{n_0},$$

and on $\mathcal{E}_{n_0,\beta}$,

$$\widehat{F}_n(q+h) \ge p + \frac{2}{n_0}.$$

On the other hand,

$$\frac{r_n}{n} \le \frac{(n+1)p+1}{n} = p + \frac{p+1}{n} \le p + \frac{2}{n} \le p + \frac{2}{n_0}.$$

Therefore $\widehat{F}_n(q+h) \ge r_n/n$, which implies

$$\hat{\tau}_\alpha \le q + h.$$

Similarly,

$$F_R^A(q-h) \le p - \eta_{n_0,\beta} - \frac{2}{n_0},$$

and hence

$$\widehat{F}_n(q-h) \le p - \frac{2}{n_0} < \frac{r_n}{n}.$$

Thus $q-h$ lies strictly below the r_n-th retained order statistic, and

$$\hat{\tau}_\alpha \ge q - h.$$

Combining the upper and lower inequalities yields

$$|\hat{\tau}_\alpha - q_P(1-\alpha)| \le \frac{\Gamma_{n_0,\beta}}{\lambda_{\min}}$$

on $\mathcal{E}_{n_0,\beta}$, which proves (14). Under the additional residual-score assumption $A(x,y) = |y - \hat{f}(x)|$, the prediction set is the symmetric residual interval displayed in the proposition. Its width is $2\hat{\tau}_\alpha$, so the stated width bound follows immediately. □

Proof. Let

$$H := \mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big).$$

Since $0 \le H \le 1$, and on $\mathcal{E}$ one has

$$H \ge 1 - \alpha - L \ge 1 - \alpha - L_0,$$

it follows that

$$H\mathbf{1}_{\mathcal{E}} \ge [1 - \alpha - L_0]_+ \mathbf{1}_{\mathcal{E}}.$$

Taking expectations gives

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}})\big) = \mathbb{E}[H] \ge [1-\alpha-L_0]_+ \mathbb{P}(\mathcal{E}) \ge (1-\beta)[1-\alpha-L_0]_+,$$

proving (15). The additive form follows from the elementary inequality

$$(1-\beta)c \ge c - \beta \qquad \text{for } 0 \le c \le 1,$$

applied with $c = [1-\alpha-L_0]_+$, and from

$$[1-\alpha-L_0]_+ \ge 1 - \alpha - L_0.$$

The final statement is the union bound applied to $\mathcal{E}_s^c \cup \mathcal{E}_{\mathrm{ref}}(\delta_\ell)^c$. □

F.12 Proof of Proposition 8

Proof. The direct bound (17) follows from Proposition 1 applied with target law $P_0 = P^\star$, together with non-negativity of coverage for the positive-part form. The target-law exact identity follows from the proof of Theorem 1 with $P_0 = P^\star$.

For the clean-component reduction, write the mixture-test coverage as

$$\begin{aligned}&\mathbb{P}_{Z_{\mathrm{new}}\sim P^\star}\{Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \mid \mathcal{G}\}\\ &\quad = (1-\varepsilon)\mathbb{P}_{Z_{\mathrm{new}}\sim P}\{Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \mid \mathcal{G}\} + \varepsilon\mathbb{P}_{Z_{\mathrm{new}}\sim Q}\{Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \mid \mathcal{G}\}.\end{aligned}$$

Dropping the non-negative dirty-component coverage gives (18). Applying Theorem 1 to the clean component gives (19). Since $D_Q \le 1$, (20) follows. □

F.13 Supplementary corollary for contaminated-mixture coverage

Corollary 4 (Conservative additive-form mixture coverage bound)**.** *Under the assumptions of Proposition 8, on any realisation of $\mathcal{G}$ satisfying those assumptions,*

$$\mathbb{P}_{Z_{\mathrm{new}}\sim P^\star}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) \ge (1-\varepsilon)(1-\alpha) - (1-\varepsilon)(1-\tilde{\varepsilon}_\star)\Delta^A_{\mathrm{trim}} - (1-\varepsilon)\tilde{\varepsilon}_\star D_Q. \tag{44}$$

In particular, since $D_Q \le 1$, $1-\varepsilon \le 1$, and $1-\tilde{\varepsilon}_\star \le 1$, the weaker but simpler bound

$$\mathbb{P}_{Z_{\mathrm{new}}\sim P^\star}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) \ge (1-\varepsilon)(1-\alpha) - \Delta^A_{\mathrm{trim}} - \tilde{\varepsilon}_\star \tag{45}$$

also holds.

Proof. Proposition 8 gives

$$\mathbb{P}_{Z_{\mathrm{new}}\sim P^\star}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) \ge (1-\varepsilon)\big[1-\alpha-\delta^{\mathrm{ref}}_{\mathrm{mix}}(D_Q)\big].$$

Expanding

$$\delta^{\mathrm{ref}}_{\mathrm{mix}}(D_Q) = (1-\tilde{\varepsilon}_\star)\Delta^A_{\mathrm{trim}} + \tilde{\varepsilon}_\star D_Q$$

gives (44).

The simple bound follows from the non-negativity of Δ^A_{trim}, $\tilde{\varepsilon}_\star$, and D_Q, together with

$$0 \le 1-\varepsilon \le 1, \qquad 0 \le 1-\tilde{\varepsilon}_\star \le 1, \qquad 0 \le D_Q \le 1.$$

□

Remark 12 (Interpreting the conservative clean-component mixture bound). Proposition 8 contains two distinct statements for mixture-test coverage. The direct retained-to-$P^\star$ transfer is the relevant scalar CDF bound when the deployment target is the contaminated mixture itself.

The clean-component reduction is more conservative. It lower-bounds mixture-test coverage using only the clean part of the mixture. It also uses only the trivial fact that dirty-test coverage is non-negative. In conditional form, the clean-component reduction gives

$$\mathbb{P}_{Z_{\mathrm{new}}\sim P^\star}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) \ge (1-\varepsilon)\big[1-\alpha-\delta^{\mathrm{ref}}_{\mathrm{mix}}(D_Q)\big]_+.$$

Expanding the retained-mixture loss yields

$$(1-\varepsilon)\Big[1-\alpha-(1-\tilde{\varepsilon}_\star)\Delta^A_{\mathrm{trim}} - \tilde{\varepsilon}_\star D_Q\Big]_+.$$

This display separates three mechanisms. First, the prefactor $(1-\varepsilon)$ is a worst-case deployed-mixture penalty. It appears because we make no assumption about coverage under an arbitrary dirty test law.

Second, the retained mixture coefficient $\tilde{\varepsilon}_\star$ is the dirty-component weight actually seen by the retained conformal quantile. Third, the clean distortion Δ^A_{trim} is the price paid for conditioning clean calibration points on the retained event.

Thus, the raw contamination level ε is not, by itself, the right diagnostic for the conformal calibration step after trimming. The relevant population tuple is

$$\big(p_c, p_d, \tilde{\varepsilon}_\star, \Delta^A_{\mathrm{trim}}, D_Q\big).$$

No additional $m^{-1/2}$ term appears in this population theorem. Such terms arise only from separate empirical certificates, or from events used to justify data-driven threshold selection.

Remark 13 (Interpretation of the $(1-\varepsilon)$ factor in the mixture-coverage bound). The prefactor $(1-\varepsilon)$ in the conservative clean-component bounds of Proposition 8 and Corollary 4 should be viewed as a worst-case contamination benchmark. It appears because the deployed-mixture coverage lower bound uses only the trivial inequality

$$\mathbb{P}_{Z_{\mathrm{new}}\sim Q}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}})\big) \geq 0.$$

Thus, the bound does not exploit any favourable structure of the contaminating law Q. Consequently, the resulting mixture-law guarantee is conservative whenever the prediction set also achieves none trivial coverage under Q. Stronger deployed-mixture coverage guarantees are possible in settings where we can separately lower-bound

$$\mathbb{P}_{Z_{\mathrm{new}}\sim Q}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}})\big).$$

They are also possible when one can exploit overlap or regularity shared by P and Q. The direct retained-to-$P^\star$ bound in Proposition 8 is the appropriate scalar transfer bound for mixture testing. The clean-component corollaries should be interpreted differently. They provide robust worst-case benchmarks under arbitrary dirty test behaviour. They should not be read as sharp descriptions of mixture-law coverage in favourable contamination regimes.

G Benchmark regimes

G.1 Ideal separation

Theorem 4 (Best-case validity under ideal separation for fixed-threshold trimming)**.** *Assume the setup of Theorem 1. Let*

$$\mathcal{A}_{\mathrm{clean}}$$

denote the realised clean auxiliary stage. This includes all auxiliary clean randomness used to construct the fitted functions A and S, as well as the fixed threshold $t^\star$.

Throughout this theorem, all conditional probability statements are interpreted given the realised clean auxiliary stage. Thus, conditional on $\mathcal{A}_{\mathrm{clean}}$, the functions A, S, and the threshold $t^\star$ are deterministic.

Suppose that, for the realised score function S and threshold $t^\star$, the fixed threshold perfectly separates the clean and contaminating distributions:

$$\mathbb{P}_{Z\sim P}\big(S(Z) \leq t^\star\big) = 1, \qquad \mathbb{P}_{Z\sim Q}\big(S(Z) \leq t^\star\big) = 0, \tag{46}$$

Here, the probabilities are taken over an independent draw Z from P or Q. The score function S and the threshold $t^\star$ are held fixed at their realised auxiliary-stage values.

Then, for every integer n with positive conditional probability under the realised auxiliary stage, conditional on $\mathcal{A}_{\mathrm{clean}}$ and on $N_{\mathrm{keep}} = n$,

$$\mathbb{P}\Big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\Big|\, \mathcal{A}_{\mathrm{clean}},\, N_{\mathrm{keep}} = n\Big) \geq \frac{\lceil (n+1)(1-\alpha)\rceil}{n+1} \geq 1-\alpha. \tag{47}$$

Consequently,

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}})\big) \geq 1-\alpha.$$

Remark 14 (Rare degenerate case under ideal separation). Under the ideal-separation assumptions,

$$\mathcal{I}_{\mathrm{keep}} = \mathcal{C} \qquad \text{and hence} \qquad N_{\mathrm{keep}} = |\mathcal{C}|$$

almost surely, where $\mathcal{C}$ is the latent clean index set in the mixture construction. Therefore

$$N_{\mathrm{keep}} \sim \mathrm{Binomial}(m, 1-\varepsilon), \qquad \mathbb{P}\big(N_{\mathrm{keep}} = 0\big) = \varepsilon^m.$$

Thus, the case $N_{\mathrm{keep}} = 0$ is not impossible. However, it is a rare corner event when $\varepsilon < 1$ and m are moderate. On this event, the algorithm returns

$$C_\alpha(x) = \mathcal{Y}.$$

This gives coverage one, but it provides no efficiency.

Proof. Work on the expanded mixture probability space with latent component labels. Let

$$\mathcal{C} := \{i \in \{1, \dots, m\} : Z_i \text{ is drawn from } P\}, \qquad \mathcal{D} := \{i \in \{1, \dots, m\} : Z_i \text{ is drawn from } Q\}.$$

Thus, $\mathcal{C}$ and $\mathcal{D}$ form a partition of $\{1, \dots, m\}$. The partition is determined by whether each calibration point comes from the clean component or the contaminating component of the mixture.

We first show that the retained set equals the clean component almost surely. Conditional on $\mathcal{A}_{\text{clean}}$, the score S and threshold $t^\star$ are fixed. By the first part of (46), for each clean calibration point $i \in \mathcal{C}$,

$$\mathbb{P}\big(S(Z_i) > t^\star \,\big|\, i \in \mathcal{C}, \mathcal{A}_{\text{clean}}\big) = 0.$$

Since the calibration sample is finite, a union bound over the clean indices gives

$$\mathbb{P}\Big(\exists i \in \mathcal{C} :\ S(Z_i) > t^\star \,\Big|\, \mathcal{C}, \mathcal{D}, \mathcal{A}_{\text{clean}}\Big) \le \sum_{i \in \mathcal{C}} \mathbb{P}\big(S(Z_i) > t^\star \,\big|\, i \in \mathcal{C}, \mathcal{A}_{\text{clean}}\big) = 0.$$

Similarly, by the second part of (46), for each dirty calibration point $i \in \mathcal{D}$,

$$\mathbb{P}\big(S(Z_i) \le t^\star \,\big|\, i \in \mathcal{D}, \mathcal{A}_{\text{clean}}\big) = 0,$$

and another finite union bound yields

$$\mathbb{P}\Big(\exists i \in \mathcal{D} :\ S(Z_i) \le t^\star \,\Big|\, \mathcal{C}, \mathcal{D}, \mathcal{A}_{\text{clean}}\Big) = 0.$$

Therefore, almost surely,

$$S(Z_i) \le t^\star \quad \text{for every } i \in \mathcal{C}, \qquad S(Z_i) > t^\star \quad \text{for every } i \in \mathcal{D}.$$

Since

$$\mathcal{I}_{\text{keep}} = \{i \in \{1, \dots, m\} : S(Z_i) \le t^\star\},$$

it follows that, almost surely,

$$\mathcal{I}_{\text{keep}} = \mathcal{C}. \tag{48}$$

Consequently,

$$N_{\text{keep}} = |\mathcal{I}_{\text{keep}}| = |\mathcal{C}|. \tag{49}$$

Next, condition on $\mathcal{A}_{\text{clean}}$. Once the full clean auxiliary stage is fixed, the fitted functions A and S and the threshold $t^\star$ are deterministic measurable objects. Conditional on the latent clean index set $\mathcal{C}$, the clean calibration points

$$\{Z_i : i \in \mathcal{C}\}$$

are independent draws from P. Since the identities (48) and (49) hold almost surely, the retained calibration scores satisfy

$$\{A_i : i \in \mathcal{I}_{\text{keep}}\} = \{A(Z_i) : i \in \mathcal{C}\}.$$

We now justify the conditioning on N_{keep} explicitly. Fix an integer $n \in \{0, \dots, m\}$. On the ideal-separation event, $N_{\text{keep}} = |\mathcal{C}|$. For every deterministic subset $I \subseteq \{1, \dots, m\}$ with $|I| = n$, conditional on

$$\mathcal{A}_{\text{clean}} \quad \text{and} \quad \mathcal{C} = I,$$

the collection

$$\{Z_i : i \in I\}$$

consists of n independent draws from P. Moreover, the clean test point

$$Z_{\text{new}} = (X_{\text{new}}, Y_{\text{new}}) \sim P$$

is independent of the calibration sample and of the clean auxiliary stage. Hence, conditional on

$$\mathcal{A}_{\text{clean}} \quad \text{and} \quad \mathcal{C} = I,$$

the nonconformity scores

$$\{A(Z_i) : i \in I\} \cup \{A(Z_{\text{new}})\}$$

are exchangeable.

Averaging over all deterministic subsets I of size n preserves this exchangeable joint law. Indeed, for each such I, the conditional law of the retained clean scores and the test score is the same. It is the law induced by $n + 1$ i.i.d. draws from P, after applying the fixed map A. Therefore, conditional on

$$\mathcal{A}_{\text{clean}} \quad \text{and} \quad N_{\text{keep}} = n,$$

the collection

$$\{A_i : i \in \mathcal{I}_{\text{keep}}\} \cup \{A(Z_{\text{new}})\}$$

is exchangeable.

We now fix the realised value $N_{\text{keep}} = n$ and consider two cases.

Case 1: $N_{\mathrm{keep}} = 0$**.** By the algorithmic definition of the procedure, $\hat{\tau}_\alpha = +\infty$. Hence

$$C_\alpha(x) = \mathcal{Y} \qquad \text{for every } x \in \mathcal{X},$$

and therefore

$$\mathbb{P}\Big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\Big|\, \mathcal{A}_{\mathrm{clean}},\, N_{\mathrm{keep}} = 0\Big) = 1.$$

Since

$$\frac{\lceil (0+1)(1-\alpha) \rceil}{0+1} = \lceil 1-\alpha \rceil = 1,$$

the bound (47) holds in this case.

Case 2: $n \geq 1$**.** Let

$$r_{\mathrm{keep}} := \big\lceil (n+1)(1-\alpha) \big\rceil.$$

If $r_{\mathrm{keep}} = n+1$, then by the algorithmic definition $\hat{\tau}_\alpha = +\infty$, and hence

$$\mathbb{P}\Big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\Big|\, \mathcal{A}_{\mathrm{clean}},\, N_{\mathrm{keep}} = n\Big) = 1 \geq \frac{r_{\mathrm{keep}}}{n+1}.$$

If instead $r_{\mathrm{keep}} \leq n$, then $\hat{\tau}_\alpha$ is the r_{keep}-th order statistic of the retained scores. By exchangeability, the usual split-conformal rank argument gives

$$\mathbb{P}\Big(A(Z_{\mathrm{new}}) \leq \hat{\tau}_\alpha \,\Big|\, \mathcal{A}_{\mathrm{clean}},\, N_{\mathrm{keep}} = n\Big) \geq \frac{r_{\mathrm{keep}}}{n+1} \geq 1-\alpha.$$

Since $Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}})$ is equivalent to $A(Z_{\mathrm{new}}) \leq \hat{\tau}_\alpha$, this proves (47) in the second case as well.

Finally, averaging (47) over the distribution of N_{keep} conditional on $\mathcal{A}_{\mathrm{clean}}$, and then over $\mathcal{A}_{\mathrm{clean}}$, yields the unconditional lower bound

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}})\big) \geq 1-\alpha.$$

□

G.2 No separation

Theorem 5 (No first-order retained-contamination gain under no separation)**.** *Work conditionally on the auxiliary sigma-field $\mathcal{G}$. Assume the setup of Theorem 1, so that A, S, and $t^\star$ are fixed after conditioning and all quantities below are $\mathcal{G}$-measurable. Suppose that the anomaly score is uninformative at the fixed threshold $t^\star$, in the retained-probability sense that*

$$p_c = \mathbb{P}_{Z\sim P}\big(S(Z) \leq t^\star\big) = \mathbb{P}_{Z\sim Q}\big(S(Z) \leq t^\star\big) = p_d =: r \in (0,1].$$

Under $Z \sim P$, define

$$K := \mathbf{1}\{S(Z) \leq t^\star\}.$$

Since $p_c = r$, $\mathbb{E}_P[K] = r$. Define the clean retained-event supremum covariance discrepancy

$$\rho_{A,K} := \sup_{t\in\mathbb{R}} \big|\mathrm{Cov}_P\big(\mathbf{1}\{A(Z) \leq t\}, K\big)\big|.$$

Also define the retained dirty nonconformity discrepancies

$$Q_{\mathrm{keep}} := Q(\,\cdot \mid S(Z) \leq t^\star), \qquad D_{Q,+} := \sup_{t\in\mathbb{R}} \big(F^A_{Q_{\mathrm{keep}}}(t) - F^A_P(t)\big)_+,$$

and

$$D_Q := \sup_{t\in\mathbb{R}} \big|F^A_{Q_{\mathrm{keep}}}(t) - F^A_P(t)\big|.$$

Then the following statements hold:

1. *The retained mass satisfies*

$$\mu_{\mathrm{keep}} = (1-\varepsilon)p_c + \varepsilon p_d = r.$$

2. *The retained mixture coefficient is unchanged:*

$$\tilde{\varepsilon}_\star = \frac{\varepsilon p_d}{(1-\varepsilon)p_c + \varepsilon p_d} = \varepsilon.$$

3. *The clean trimming distortion is exactly controlled by the retained-event supremum covariance discrepancy:*

$$\Delta^A_{\mathrm{trim}} = \frac{\rho_{A,K}}{r}.$$

In particular, if $\rho_{A,K} \le c$, then

$$\Delta^A_{\mathrm{trim}} \le \frac{c}{r}.$$

As a strong sufficient special case, if $S(Z)$ and $A(Z)$ are independent under P, then $\rho_{A,K} = 0$ and hence

$$\Delta^A_{\mathrm{trim}} = 0.$$

More generally, the relevant requirement is not literal independence but small $\rho_{A,K}$.

4. *Consequently, the mixture upper-bound consequence of Theorem 1 reduces to*

$$\begin{aligned}\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) &\ge \left[1 - \alpha - (1-\varepsilon)\frac{\rho_{A,K}}{r} - \varepsilon D_{Q,+}\right]_+ \\ &\ge \left[1 - \alpha - (1-\varepsilon)\frac{\rho_{A,K}}{r} - \varepsilon D_Q\right]_+.\end{aligned}$$

Since $D_Q \le 1$, the worst-case dirty-side version is

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) \ge \left[1 - \alpha - \varepsilon - (1-\varepsilon)\frac{\rho_{A,K}}{r}\right]_+.$$

In particular, if $\rho_{A,K} \le c$, then

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) \ge \left[1 - \alpha - \varepsilon - (1-\varepsilon)\frac{c}{r}\right]_+.$$

In the independence special case this further simplifies to

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) \ge [1 - \alpha - \varepsilon]_+.$$

Thus, in the no-separation regime $p_c = p_d = r$, fixed-threshold trimming does not reduce the retained mixture coefficient. In this case, $\tilde{\varepsilon}_\star$ remains equal to ε. Any further benefit must come from one of two sources. The first is structural information implying $D_{Q,+} < 1$. The second is that the retained event is nearly harmless for the clean nonconformity score. Equivalently, $\rho_{A,K}$ must be small.

Remark 15 (Interpretation of the no-separation regime). Theorem 5 gives a negative characterisation of the retained-contamination part of the mixture upper-bound diagnostic. It is not a universal impossibility theorem.

When $p_c = p_d = r$, trimming does not reduce the retained mixture coefficient:

$$\tilde{\varepsilon}_\star = \varepsilon.$$

In the endpoint case $r = 1$, no observations are trimmed. Then $K = 1$ P-almost surely, $\rho_{A,K} = 0$, and

$$\Delta^A_{\mathrm{trim}} = 0.$$

However, the retained mixture coefficient remains

$$\tilde{\varepsilon}_\star = \varepsilon.$$

For $r < 1$, trimming may also distort the clean nonconformity distribution. This distortion is measured exactly by

$$\Delta^A_{\mathrm{trim}} = \frac{\rho_{A,K}}{r}.$$

Thus, without retained-probability separation, trimming is not a free lunch. It leaves the retained mixture coefficient unchanged. It can also introduce a covariance-distortion cost.

The refined dirty-side factor $D_{Q,+}$ identifies the only way the retained dirty law can help in this regime. Its retained nonconformity distribution must itself be close to the clean nonconformity distribution. Therefore, the theorem rules out improvement through the retained contamination coefficient. It does not rule out improvement through a favourable change in the retained dirty score distribution Q_{keep}.

Proof. By assumption $p_c = p_d = r$, hence

$$\mu_{\mathrm{keep}} = (1-\varepsilon)p_c + \varepsilon p_d = (1-\varepsilon)r + \varepsilon r = r,$$

proving part (1).

The retained mixture coefficient is

$$\tilde{\varepsilon}_\star = \frac{\varepsilon p_d}{(1-\varepsilon)p_c + \varepsilon p_d}.$$

Substituting $p_c = p_d = r$ gives

$$\tilde{\varepsilon}_\star = \frac{\varepsilon r}{(1-\varepsilon)r + \varepsilon r} = \varepsilon,$$

proving part (2).

By Proposition 9 applied with $p_c = r$, we have the exact identity

$$\Delta^A_{\mathrm{trim}} = \frac{1}{r}\sup_{t\in\mathbb{R}}\left|\mathrm{Cov}_P\big(\mathbf{1}\{A(Z)\le t\}, K\big)\right| = \frac{\rho_{A,K}}{r}.$$

This proves part (3). The bound with $\rho_{A,K} \le c$ is immediate. If $S(Z)$ and $A(Z)$ are independent under P, then $K = \mathbf{1}\{S(Z) \le t^\star\}$ is independent of each event $\{A(Z) \le t\}$, so all the displayed covariances vanish and $\rho_{A,K} = 0$.

Finally, substitute

$$\tilde{\varepsilon}_\star = \varepsilon, \qquad \Delta^A_{\mathrm{trim}} = \frac{\rho_{A,K}}{r}$$

into the one-sided mixture upper-bound consequence of Theorem 1. Since $\Delta^A_{\mathrm{trim},+} \le \Delta^A_{\mathrm{trim}} = \rho_{A,K}/r$, this gives the refined bound

$$\mathbb{P}\big(Y_{\mathrm{new}} \in C_\alpha(X_{\mathrm{new}}) \,\big|\, \mathcal{G}\big) \ge \left[1 - \alpha - (1-\varepsilon)\frac{\rho_{A,K}}{r} - \varepsilon D_{Q,+}\right]_+.$$

The displayed bound with D_Q follows from $D_{Q,+} \le D_Q$. Since $D_Q \le 1$, the worst-case dirty-side version and its consequences follow. □

H Detailed experimental summaries

H.1 Experimental setup

The empirical study is organised around the clean-test retained-law identity and mixture upper-bound diagnostic in Theorem 1, the contaminated-mixture comparison in Proposition 8, and the ideal- and no-separation benchmark regimes. All simulation summaries use 100 Monte Carlo repetitions. Code to reproduce our experiments is available via this GitHub link.

In the regression experiments, the conformal nonconformity score is the absolute residual from a linear predictive backbone. A clean fitting split trains the prediction rule and, when needed, an auxiliary anomaly-score model. An independent clean reference split is used only when the threshold policy requires clean reference threshold selection. The final contaminated calibration sample is then trimmed by the fixed anomaly score and conformalised using the retained calibration scores.

Several experiments use a plug-in Stein score-norm as the anomaly score, following the kernelised Stein discrepancy and related kernel-discrepancy literature [Gorham and Mackey, 2015, Liu et al., 2016, Gorham and Mackey, 2017, Jitkrittum et al., 2020, Oates, 2022, Wang et al., 2023, Liu et al.,

2026]. For a representation $u = T(x)$, a fitted clean score model $\hat{s}$, and a smooth scalar kernel k, define

$$S(z) = \sqrt{\max\{h_{\hat{s}}(T(x), T(x)), 0\}},$$

where

$$h_{\hat{s}}(u, v) = \hat{s}(u)^\top k(u, v)\hat{s}(v) + \hat{s}(u)^\top \nabla_v k(u, v) + \hat{s}(v)^\top \nabla_u k(u, v) + \mathrm{tr}\{\nabla_u \nabla_v^\top k(u, v)\}.$$

The one-dimensional Gaussian experiments use $T(x) = x$. A Gaussian clean score is fitted on the independent clean fitting split

$$\hat{s}(u) = -\frac{u - \hat{\mu}_X}{\hat{\sigma}_X^2},$$

where $\hat{\mu}_X$ and $\hat{\sigma}_X^2$ are the empirical means and variances of the clean fitting covariates. The kernel is the Gaussian RBF kernel

$$k(u, v) = \exp\{-(u - v)^2/(2h^2)\},$$

with bandwidth h chosen by the median heuristic on the transformed clean fitting split.

Thresholds labelled "Stein q" use the corresponding clean-reference or population q-quantile of this anomaly score, as stated in the table caption. In multivariate synthetic geometry checks, the same formula is applied componentwise with $T(x) = x$ and the empirical covariance-scaled Gaussian score. Rows labelled Mahalanobis or GMM Mahalanobis use the stated scores instead of the Stein score-norm.

These rows should be read as one covariate-geometric score implementation. The evaluated coverage statements are the score-agnostic retained-law bounds, not a separate Stein-specific certificate.

The run summaries distinguish *mixture diagnostics* from *finite/audit-style lower bounds*. The former use knowledge of the simulation laws to evaluate $p_c, p_d, \tilde{\varepsilon}_\star, \Delta^A_{\mathrm{trim},+}, D_{Q,+}$, and $L_{\mathrm{mix},+}$. The latter use independent structure/audit samples and concentration corrections. They are reported as conservative certification diagnostics, not as the exact/scalar theorem bound. In the displayed experiments, these lower bounds are finite and directionally informative, but remain too conservative to certify nominal 0.9 coverage.

H.2 Separation phase diagram

Table 5 shows a representative slice of the separation phase diagram at $\varepsilon = 0.3$ and $m = 800$. When separation is small, the selected method is either ordinary split or a clean-reference trimming rule that retains almost all dirty points. As separation increases, p_d and $\tilde{\varepsilon}_\star$ fall, and empirical clean coverage moves into the near-nominal range.

Separation	Selected rule	Clean coverage	Width	p_d	$\tilde{\varepsilon}_\star$	$L_{\mathrm{mix},+}$
0.0	Ordinary split	$0.8593\,[0.8557, 0.8629]$	1.7720	1.0000	0.3000	0.6555
0.5	Stein clean-ref $q = 0.950$	$0.8595\,[0.8558, 0.8631]$	1.7717	0.9601	0.3035	0.6498
1.0	Stein clean-ref $q = 0.950$	$0.8635\,[0.8599, 0.8670]$	1.7907	0.8753	0.2841	0.6662
2.0	Stein clean-ref $q = 0.950$	$0.8806\,[0.8775, 0.8836]$	1.8731	0.4706	0.1749	0.7550
4.0	Stein clean-ref $q = 0.950$	$0.9013\,[0.8987, 0.9039]$	1.9888	0.0064	0.0028	0.8934
8.0	Stein clean-ref $q = 0.950$	$0.9015\,[0.8988, 0.9041]$	1.9900	0.0000	0.0000	0.8958

Table 5: Separation phase diagram slice at $\varepsilon = 0.3$ and $m = 800$. The transition is driven by dirty retention p_d and retained mixture coefficient $\tilde{\varepsilon}_\star$, not by trimming alone.

H.3 Clean-reference and threshold-policy diagnostics

Clean-reference $q = 0.950$ thresholding is efficient but can be coverage-aggressive. In the clean-reference concentration experiment, increasing the clean reference size from 64 to 512 changes coverage only from 0.8847 to 0.8872. The dominant effect is not reference-sample noise, but the population clean distortion induced by $q = 0.950$. Table 6 summarises the threshold-policy comparison. The KSD-geometric policy is more conservative and closer to nominal coverage than fixed or clean-reference $q = 0.950$, but gives wider intervals.

Policy	Clean coverage	Width	p_c	Δ_+	$L_{\mathrm{mix},+}$	Audit LB$_+$
Clean-reference $q = 0.950$	$0.8923\,[0.8880, 0.8967]$	2.9203	0.9453	0.0153	0.8847	0.6031
Fixed $q = 0.950$	$0.8927\,[0.8883, 0.8970]$	2.9231	0.9486	0.0145	0.8854	0.5996
KSD-geometric	$0.8979\,[0.8937, 0.9020]$	2.9741	0.9861	0.0048	0.8951	0.6138
Oracle selected	$0.8927\,[0.8883, 0.8970]$	2.9231	0.9486	0.0145	0.8854	0.5996

Table 6: Threshold-policy comparison. The KSD-geometric policy improves the coverage-efficiency trade-off relative to $q = 0.950$ rules. However, the audit-style lower bound remains far below nominal coverage.

Localised threshold policies must also report fallback rates. In the localised certificate experiment, loc4/R30 gives coverage 0.8998 with fallback rate 0.01. By contrast, loc6/R30 gives similar coverage, 0.8977, but with fallback rate 0.96. The latter should be read as a fallback-driven policy outcome, not as stable success of the primary localised certificate.

H.4 Large-calibration behaviour

Table 7 studies the nominal $q = 0.950$ policy as the calibration size grows, using a separate large-calibration run. This run refits the auxiliary score and threshold policy. It is therefore not the same population-threshold instance as Table 1. Dirty-retention numbers should not be compared as if p_d were fixed across the two tables.

Within the large-calibration run, the one-sided mixture diagnostic is approximately $L_{\mathrm{mix},+} = 0.8853$. The conservative lower bound rises from 0.7132 at $m = 320$ to 0.8464 at $m = 10000$. This supports the interpretation that finite-sample slack shrinks with calibration size, while the limiting clean-coverage level is governed by retained-law-to-clean-law transfer.

m	Clean coverage	Width	$L_{\mathrm{mix},+}$	Cons. LB$_+$	Audit LB$_+$
320	$0.8878\,[0.8838, 0.8918]$	2.8659	0.8853	0.7132	0.6560
1000	$0.8903\,[0.8880, 0.8925]$	2.8775	0.8853	0.7788	0.7225
3000	$0.8909\,[0.8894, 0.8924]$	2.8812	0.8853	0.8191	0.7633
10000	$0.8888\,[0.8878, 0.8898]$	2.8601	0.8853	0.8464	0.7910

Table 7: Large-calibration behaviour for the nominal Stein $q = 0.950$ policy in a separate large-calibration run. Increasing m reduces the slack in the conservative lower bound, but the population mixture diagnostic remains below the nominal 0.9 level.

H.5 Label-only and stress-test regimes

Label-only contamination has $Q_X = P_X$, so a covariate-only anomaly score has no direct signal for corrupted labels. In the label-only experiment, ordinary split conformal has clean coverage 0.9723, 0.9998, and 1.0000 at $\varepsilon = 0.1, 0.2, 0.3$, with widths 2.7200, 5.6192, and 7.8897, respectively. The issue is therefore not undercoverage, but inefficient overcoverage caused by response-side corruption. This lies outside the favourable score-visible trimming regime.

The stress tests separate the mechanisms. Heavy-tail stress gives Stein clean-reference coverage 0.8978 with width 2.9616, suggesting useful trimming. High-leverage stress causes severe overcoverage and very wide intervals for non-trimming baselines. Thus, selecting maximal coverage alone can choose inefficient intervals. Response stress has $p_d \approx p_c$, so covariate anomaly trimming has little leverage over response-side corruption.

H.6 Non-Gaussian score geometry

Table 8 summarises representative non-Gaussian geometry results. The selected score differs across regimes. This is expected under the score-agnostic theorem. Validity depends on the retained law and score CDFs after the score is fixed. Efficiency and diagnostic quality depend on how well the score matches the clean geometry and the direction of contamination.

Setting	Selected score	Clean coverage	Width	p_d	$\tilde{\varepsilon}_\star$	$L_{\mathrm{mix},+}$
Bimodal moderate	Mahalanobis	$0.8918\ [0.8876, 0.8960]$	1.9351	0.4401	0.1360	0.7822
Bimodal overlap	GMM Mahalanobis	$0.8948\ [0.8904, 0.8991]$	1.9516	0.3590	0.1144	0.7999
Bimodal separated	Stein score-norm	$0.8997\ [0.8962, 0.9031]$	1.9765	0.1985	0.0497	0.8571
Banana bridge	GMM Mahalanobis	$0.8961\ [0.8934, 0.8989]$	1.6758	3.4×10^{-5}	9.0×10^{-6}	0.8921
Two moons gap	Stein score-norm	$0.9708\ [0.9654, 0.9761]$	2.3689	0.1519	0.0777	0.8787

Table 8: Non-Gaussian score-geometry diagnostics. The Stein score-norm is useful in some settings. Mixture-aware Mahalanobis scores are competitive or better in others. Coverage-best or overcoverage rows should not be interpreted as the shortest valid methods.

H.7 Compute resources

The experiments do not require specialised compute. All reported simulations were run on a personal laptop using Python 3.12, an AMD Ryzen 9 8945HS CPU with Radeon 780M Graphics, and 32GB memory. Reproducing the full set of reported experiments required approximately 40 minutes of wall-clock time. No GPU acceleration was used.

H.8 Semi-synthetic ACS classification

We evaluate the retained law diagnostic on the 2018 American Community Survey (ACS) ACSEmployment classification task using Folktables [Ding et al., 2021]. California (CA) defines the clean target distribution P. Texas (TX), New York (NY), Florida (FL), and their equal mixture define the contaminating distribution Q. We consider contamination levels $\varepsilon \in \{0.1, 0.2, 0.3\}$.

All methods use the same logistic regression predictor and randomised adaptive prediction sets (APS) score at nominal coverage 0.9. The experiment uses 50 Monte Carlo repetitions. We report the results for the equal mixture of TX, NY, and FL under the fixed 0.95 threshold policy.

We compare direct clean calibration, ordinary APS, source oracle trimming, and three covariate based trimming methods: GMM Mahalanobis, Isolation Forest, and Mahalanobis. Source labels are used only to construct the contamination and evaluate the oracle. They are not available to the three implementable trimming methods.

Results and interpretation. Table 9 shows that all three practical anomaly scores have similar clean and dirty retention probabilities. Across these methods, p_c ranges from 0.9490 to 0.9502, while p_d ranges from 0.9477 to 0.9527. Consequently, the retained contamination coefficient $\tilde{\varepsilon}$ remains close to the original contamination level ε.

The practical trimming methods provide little improvement over ordinary APS. For $\varepsilon = 0.1$, their mean clean coverage ranges from 0.9003 to 0.9006, compared with 0.9004 for ordinary APS. For $\varepsilon = 0.2$, the corresponding range is 0.9003 to 0.9004, compared with 0.9006 for ordinary APS. For $\varepsilon = 0.3$, it is 0.8989 to 0.8992, compared with 0.8992 for ordinary APS.

The largest absolute difference in mean coverage between a practical trimming method and ordinary APS is only 0.0003 in this experiment. Mean prediction set sizes also remain nearly unchanged.

Source oracle trimming removes all contaminating observations while retaining all clean observations, giving $p_c = 1$, $p_d = 0$, and $\tilde{\varepsilon} = 0$. Its mean coverage remains close to nominal coverage. This shows that even perfect identification of contaminated observations need not produce a large coverage improvement when ordinary APS already achieves near nominal coverage.

These results support the coefficient level separation dichotomy in Proposition 4. When $p_c \approx p_d$, trimming cannot substantially reduce contamination through the retained mixture coefficient. This does not rule out improvements through changes in the retained score distributions. Rather, it shows that common covariate anomaly scores provide little practical benefit in this overlapping domain setting.

ε	Method	Clean coverage [95% MC CI]	Mean set size	p_c	p_d	$\tilde{\varepsilon}$
0.1	Direct clean calibration	0.8990 [0.8974, 0.9005]	1.4386	—	—	—
	Ordinary APS	0.9004 [0.8991, 0.9018]	1.4432	1.0000	1.0000	0.1000
	Source-oracle trim	0.9009 [0.8995, 0.9024]	1.4446	1.0000	0.0000	0.0000
	GMM-Mahalanobis trim	0.9006 [0.8993, 0.9019]	1.4438	0.9494	0.9477	0.0998
	Isolation Forest trim	0.9003 [0.8990, 0.9016]	1.4430	0.9490	0.9527	0.1003
	Mahalanobis trim	0.9005 [0.8991, 0.9018]	1.4433	0.9502	0.9482	0.0998
0.2	Direct clean calibration	0.8990 [0.8974, 0.9005]	1.4386	—	—	—
	Ordinary APS	0.9006 [0.8989, 0.9022]	1.4438	1.0000	1.0000	0.2000
	Source-oracle trim	0.9015 [0.8998, 0.9033]	1.4468	1.0000	0.0000	0.0000
	GMM-Mahalanobis trim	0.9004 [0.8987, 0.9021]	1.4431	0.9494	0.9477	0.1997
	Isolation Forest trim	0.9003 [0.8986, 0.9020]	1.4429	0.9490	0.9527	0.2006
	Mahalanobis trim	0.9004 [0.8988, 0.9021]	1.4432	0.9502	0.9482	0.1997
0.3	Direct clean calibration	0.8990 [0.8974, 0.9005]	1.4386	—	—	—
	Ordinary APS	0.8992 [0.8974, 0.9010]	1.4393	1.0000	1.0000	0.3000
	Source-oracle trim	0.9002 [0.8984, 0.9019]	1.4420	1.0000	0.0000	0.0000
	GMM-Mahalanobis trim	0.8992 [0.8974, 0.9010]	1.4393	0.9494	0.9477	0.2996
	Isolation Forest trim	0.8989 [0.8971, 0.9007]	1.4385	0.9490	0.9527	0.3008
	Mahalanobis trim	0.8990 [0.8972, 0.9007]	1.4384	0.9502	0.9482	0.2996

Table 9: Semi synthetic ACS classification under cross state calibration contamination. California is the clean target, and an equal mixture of Texas, New York, and Florida is the contaminating distribution. Results are averaged over 50 Monte Carlo repetitions at nominal coverage 0.9. Brackets give 95% Monte Carlo confidence intervals for mean clean coverage. Mean set size is the average number of labels in the prediction set. Here, p_c and p_d are the clean and dirty retention probabilities, and $\tilde{\varepsilon}$ is the retained contamination coefficient. Direct clean calibration uses clean calibration data, while source oracle trimming uses true component labels. The three practical trimming methods use covariate anomaly scores without source labels.

H.9 Label only contamination and structured robust baselines

We conduct a separate label only contamination experiment for the California ACS classification task [Ding et al., 2021]. Since the contaminating mechanism keeps the covariate distribution unchanged, $Q_X = P_X$. As a result, covariate only anomaly scores cannot preferentially remove corrupted observations by separating retention probabilities.

We compare ordinary APS against Contamination Robust Conformal Prediction (CRCP) [Clarkson et al., 2024]. We also compare ordinary label conditional APS against the known noise method of Sesia et al. [2025]. The robust methods are assessed under the true synthetic noise mechanism.

Table 10 presents marginal coverage and mean prediction set size for ordinary APS and CRCP. For the label conditional methods, we present worst class coverage and mean prediction set size.

CRCP lessens the conservatism of ordinary APS. At contamination level $\varepsilon = 0.2$, for instance, its marginal coverage is 0.9074, while ordinary APS has 0.9309, and its mean set size falls from 1.5429 to 1.4662.

The method of Sesia et al. [2025] raises worst class coverage at all three evaluated noise levels. It also lowers mean prediction set size at the two higher noise levels.

These results indicate that methods built for structured label corruption can be more appropriate than covariate trimming when the contamination mechanism is known. They tackle a different problem from the retained law diagnostic, which describes the effect of a proposed trimming rule instead of directly correcting the label noise mechanism.

H.10 Allocation of a Limited Clean Data Budget

We compare three uses of a limited clean labelled sample. These uses are direct conformal calibration, independent threshold tuning, and auditing a fixed final prediction set.

ε	Ordinary APS Cov. / Size	CRCP Cov. / Size	Ordinary LC-APS Worst-Class / Size	Sesia LC Worst-Class / Size
0.1	0.9150 / 1.4905	0.9017 / 1.4478	0.9039 / 1.4867	0.9109 / 1.4864
0.2	0.9309 / 1.5429	0.9074 / 1.4662	0.9104 / 1.5434	0.9163 / 1.5086
0.3	0.9448 / 1.5912	0.9141 / 1.4877	0.9164 / 1.5987	0.9247 / 1.5371

Table 10: Classification results under label-only contamination. Ordinary APS and CRCP report marginal coverage. The label-conditional methods report worst-class coverage. Set size is reported for all methods. The robust baselines use the true synthetic noise mechanism.

Clean budget	Direct cal. coverage	Direct size	Tuning coverage	Tuning size	No-trim rate	Audit mean 95% LB	Audit cert. rate
32	0.9171	1.5099	0.8993	1.4381	1.00	0.760	0.02
64	0.9021	1.4529	0.8993	1.4381	1.00	0.820	0.04
128	0.9071	1.4650	0.8993	1.4381	1.00	0.850	0.04
256	0.9042	1.4545	0.8993	1.4381	1.00	0.865	0.10
512	0.8968	1.4310	0.8993	1.4381	1.00	0.876	0.08

Table 11: Comparison of three uses of a limited clean-data budget over 50 Monte Carlo repetitions. Direct calibration constructs a new prediction set. Independent tuning selects a policy using clean data. Auditing certifies a fixed final prediction set. The audit certification rate is the fraction of repetitions in which the exact 95% lower bound reaches 0.9.

Direct calibration uses the clean sample to construct a new conformal cut-off. Independent tuning uses the sample to select a trimming policy before final calibration. Auditing instead assesses the coverage of an already constructed prediction set without modifying that set.

We consider clean budgets of 32, 64, 128, 256, and 512 observations. All results are averaged over 50 Monte Carlo repetitions.

In this experiment, independent tuning always selects the no trimming fallback. The candidate prediction sets have clean coverage close to the nominal target, so no candidate has a sufficient margin to pass the joint tuning screen.

The fixed final prediction set has a mean clean coverage of 0.8994. However, the exact 95% binomial audit certificate verifies coverage of at least 0.9 in only 2% to 10% of the repetitions.

Direct clean calibration provides a different prediction set and the usual finite sample marginal coverage guarantee when its exchangeability assumptions hold. In contrast, auditing provides an independent confidence statement about an already selected set.

These findings clarify the different roles of calibration, tuning, and auditing. Direct clean calibration is often preferable when enough clean labelled data are available and marginal validity is the primary objective. Auditing is valuable when the final prediction set must remain fixed, but its ability to certify a target coverage level depends on the available sample size and the coverage margin.

H.11 Diagnostic ranking and coverage efficiency trade off

We ask whether the population diagnostics retain the empirical ordering of candidate trimming policies. To this end, we compare the component wise mixture lower bound L_{mix} with a direct scalar gap lower bound L_{direct}. Agreement in ranking is summarised by Kendall's τ_b and pairwise concordance. The component wise diagnostic yields Kendall's $\tau_b = 0.496$, with a stratified bootstrap 95% confidence interval of [0.207, 0.569]. The direct scalar gap diagnostic yields $\tau_b = 0.482$, with a confidence interval of [0.188, 0.569]. Across the 843 resolved comparisons, concordance rates are 96.9% and 97.3%, respectively. These resolved pairs represent approximately 58.5% of all 1,440 pairs. The remaining comparisons remain unresolved under the evaluation criterion. Taken together, these results indicate that the diagnostics can offer useful relative rankings among empirically distinguishable candidates. However, they do not establish that the component wise lower bound

Diagnostic	Kendall τ_b	95% CI	Concordant	Discordant	Unresolved	Concordance among resolved
Componentwise L_{mix}	0.496	[0.207, 0.569]	817	26	597	96.9%
Direct scalar-gap L_{direct}	0.482	[0.188, 0.569]	820	23	597	97.3%

Table 12: Ranking consistency of the component wise and direct scalar-gap diagnostics.

Method family	Number of frontier appearances
Isolation Forest trimming	55
GMM-Mahalanobis trimming	54
Mahalanobis trimming	52
Ordinary APS	12

Table 13: Pareto frontier appearances by method family.

is necessary or sufficient for actual coverage improvement. Coverage ranking, moreover, does not capture prediction set efficiency. We therefore examine empirical Pareto frontiers based on coverage and prediction set size. Isolation Forest, GMM Mahalanobis, Mahalanobis, and ordinary APS each appear on the empirical frontiers. Across the evaluated settings, no single method dominates the others. Accordingly, the diagnostics are best used to interpret and screen candidate trimming rules, whereas empirical coverage and set size remain necessary for judging practical performance.

H.12 Additional Covariance and Scalar-Bound Checks

The covariance identity is visible in the data Proposition 3 predicts that the clean trimming distortion equals a one-sided covariance under P

$$\Delta^{A}_{\mathrm{trim},+} \;=\; p_c^{-1} \sup_t \big\{ \mathrm{Cov}_P(\mathbf{1}\{A(Z) \le t\},\, K) \big\}_+ .$$

Reducing $\Delta^{A}_{\mathrm{trim},+}$ is therefore equivalent to reducing the implied covariance envelope $p_c, \Delta^{A}_{\mathrm{trim},+}$, after adjusting for the clean retention probability. Table 14 interprets three score-visible thresholds through this identity. As q increases from 0.950 to 0.990, the threshold becomes less aggressive. It retains more clean mass and trims fewer points. Dirty retention is negligible in all three rows. The implied covariance envelope decreases by about a factor of four, from 1.30×10^{-2} to 3.6×10^{-3}. The clean-side trimming cost decreases accordingly. Improvement in this regime is therefore mainly a clean score-neutrality effect, not stronger contamination removal.

The scalar bound is finite-sample sharp Theorem 2 gives the precise minimax meaning of sharpness. Once the score-law comparison is reduced to the scalar constraint $d^{A}_{R\to P,+} \le d$, the bound $L^{m,\mu}_{\mathrm{fs}}(d)$ is the best possible uniform lower bound.

Table 15 reports this curve at the main calibration scale $m = 320$, retention probability $\mu = 0.95$, and $\alpha = 0.1$. The finite-sample bound is slightly above the asymptotic line $[1 - \alpha - d]_+$. This is because the inclusive conformal quantile is finite-sample conservative. The sharpness construction in the proof attains this curve exactly.

I Additional experimental diagnostics

I.1 Reporting status

The simulation summaries report the one-sided quantities required by Theorem 1: $\Delta^{A}_{\mathrm{trim},+}$, $D_{Q,+}$ when the retained dirty component is estimable, the one-sided transfer loss, $L_{\mathrm{mix},+}$, a conservative finite/audit lower bound, and the symmetric diagnostics used in supplementary sensitivity checks.

The manuscript keeps these objects separate. Population diagnostics use knowledge of the simulation laws. Conservative lower bounds use independent audit information and concentration corrections. Empirical coverage is estimated from finite Monte Carlo test sets.

Rows with $p_d = 0$ display the retained dirty discrepancy as "N/A", rather than as a numerical failure. When p_d is extremely small, $D_{Q,+}$ is a conditional rare-event quantity and may be unstable as a

Threshold	p_c	p_d	$\Delta^A_{\mathrm{trim},+}$	$p_c\,\Delta^A_{\mathrm{trim},+}$
Stein $q = 0.950$	0.9494	2.8×10^{-5}	0.0137	0.0130
Stein $q = 0.975$	0.9752	6.1×10^{-5}	0.0075	0.0073
Stein $q = 0.990$	0.9897	2.6×10^{-4}	0.0036	0.0036

Table 14: Covariance-identity interpretation of three score-visible thresholds. The last column reports the implied one-sided covariance envelope $\sup_t\{\mathrm{Cov}_P(\mathbf{1}\{A \le t\}, K)\}_+$ up to rounding. Less aggressive trimming reduces this envelope by roughly a factor of four.

d	$L_{\mathrm{fs}}^{320,\,0.95}(d)$	$[1-\alpha-d]_+$
0.000	0.9015	0.9000
0.002	0.8995	0.8980
0.005	0.8965	0.8950
0.010	0.8915	0.8900
0.020	0.8815	0.8800
0.050	0.8515	0.8500

Table 15: Finite-sample scalar lower bound $L_{\mathrm{fs}}^{m,\mu}(d)$ and its asymptotic counterpart for $m = 320$, $\mu = 0.95$, and $\alpha = 0.1$. Theorem 2 shows that the L_{fs} column is attainable within the scalar-discrepancy class.

stand-alone diagnostic. In such rows, the dirty contribution $\tilde{\varepsilon}_\star D_{Q,+}$, or its worst-case upper bound $\tilde{\varepsilon}_\star$, is more informative for the coverage lower bound.

I.2 Population diagnostics versus conservative lower bounds

The one-sided mixture diagnostic can be close to the observed clean coverage. In the main score-visible row, the observed clean coverage is 0.89499, while $L_{\mathrm{mix},+} = 0.89638$. The Monte Carlo interval contains the mixture diagnostic, so we interpret this as agreement within simulation error.

By contrast, the conservative finite/audit lower bounds are much smaller. Thus, $L_{\mathrm{mix},+}$ explains the retained-law regime, while the conservative lower bounds are safe but usually too loose to certify nominal 0.9 coverage in these examples.

I.3 Coverage-best tables

Coverage-best tables are kept only as appendix diagnostics. They include oracle or empirical coverage-maximising choices, which often select severely over-covering and wide intervals. For example, in non-Gaussian coverage-best summaries, methods with coverage near one can have much wider intervals than fixed or independently tuned policies. These tables should not be described as implementable best-method comparisons. They are maximum empirical-coverage benchmarks.

I.4 Localised certificates

The localised-certificate summaries are aggregated at the policy level, rather than split by realised fallback labels. Reporting the fallback rate is essential. In these summaries, loc4/R30 has fallback rate 0.01 and coverage 0.89979, while loc6/R30 has fallback rate 0.96 and coverage 0.89766. The latter mainly reflects the fallback policy. It is not evidence that the primary localised certificate reliably selects the displayed threshold.

I.5 Recommended table convention

For experimental tables supporting the retained-law claim, report the method, threshold source, clean coverage with Monte Carlo interval, average width with Monte Carlo interval, p_c, p_d, $\tilde{\varepsilon}_\star$, $\Delta^A_{\mathrm{trim},+}$, $D_{Q,+}$ or the dirty contribution $\tilde{\varepsilon}_\star D_{Q,+}$, $L_{\mathrm{mix},+}$, the conservative finite/audit lower bound, and the fallback rate when the threshold policy is adaptive. This convention keeps theoretical transfer quantities, certification diagnostics, and empirical performance separate.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main contribution as a retained-law diagnostic for fixed-threshold trimming under calibration contamination, rather than as a universal validity-restoration method. The claims are matched to the theoretical results in Section 3 and to the simulation diagnostics in Section 4, with the distinction between population diagnostics and numerical certificates stated explicitly.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses the main limitations, including calibration-side i.i.d. contamination, marginal rather than conditional or subgroup coverage, the need for independent audit information for numerical certificates, and the restriction of same-sample threshold selection to finite-grid empirical-CDF bounds. Section 4 also reports failure modes, including no-separation and label-only contamination regimes.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical statements specify the conditioning, independence, retained-mass, fixed-threshold, and audit assumptions under which they hold. Proofs of the retained-law transfer, exact finite-sample identity, scalar sharpness result, same-sample selection bound, and audit certificates are provided in Appendix F, with auxiliary conventions and lemmas in Appendices C–E.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experiments are based on synthetic data-generating mechanisms, and the paper reports the clean and contaminated laws, contamination level, calibration size, nominal level, thresholding conventions, Monte Carlo repetitions, and anomaly-score construction. Appendix H provides additional implementation details for the Stein score-norm, threshold policies, and supplementary simulation regimes.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymised code package is provided with the supplemental material, including scripts for generating the synthetic data, running the conformal calibration experiments, and reproducing the main tables and appendix diagnostics. The reported datasets are synthetic and can be generated from the specifications given in the paper.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: Section 4 specifies the clean and contaminated data-generating mechanisms, contamination level, nominal miscoverage level, calibration size, Monte Carlo repetitions, and independent fitting/reference/calibration structure. Appendix H gives further details on the regression backbone, absolute-residual nonconformity score, plug-in Stein score-norm, kernel choice, bandwidth heuristic, and threshold policies.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main experimental tables report Monte Carlo intervals for empirical clean coverage, and the text states that summaries use 100 Monte Carlo repetitions. The paper also separates empirical coverage from population diagnostics and conservative audit-style lower bounds, preventing the intervals from being misinterpreted as formal coverage certificates.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are lightweight synthetic Monte Carlo simulations and can be reproduced on a personal laptop. The reported runs were executed with Python 3.12 on a machine with an AMD Ryzen 9 8945HS CPU and 32GB memory; reproducing all reported experiments took approximately 40 minutes of wall-clock time.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is methodological and uses synthetic experiments rather than human-subject data or sensitive datasets.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The work may improve the reliability of conformal prediction workflows by clarifying when trimming can and cannot help under calibration contamination.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, scraped datasets, generative models, or dual-use assets. The released material, if any, consists of code for synthetic simulations and diagnostic experiments.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The reported experiments use synthetic data generated from distributions specified in the paper, and no existing dataset or pretrained model is required for the main results. Standard prior methods and theoretical tools are credited through citations.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset is the anonymised experimental code for reproducing the synthetic simulations and diagnostic tables. The code package should include configuration files, random seeds, environment information, and instructions for reproducing the main and appendix results.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, human-subject experiments, participant instructions, or compensation.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, crowdsourcing, or personally identifiable data, so IRB approval or equivalent human-subject review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology, theory, and experiments do not use LLMs as an important, original, or non-standard component. Any use of LLMs solely for writing, editing, or formatting would not affect the scientific content and does not require declaration under this question.