REPLICABLE SPLIT-CONFORMAL CALIBRATION IN COVERAGE UNITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Two laboratories that calibrate the same prediction model on their own data deploy different prediction sets. The reason is that the calibrated cutoff is computed from continuous scores, so two calibrations never produce the same cutoff, even when both are correct on average. As a result, predictions stored by one laboratory cannot be reused by another, and an auditor cannot reproduce a vendor's prediction sets. We introduce replicable split-conformal calibration. It rounds each calibrated cutoff up to the nearest of a few values that both laboratories agree on in advance, so that both deploy identical prediction sets with a probability they choose, while each keeps its own coverage guarantee. The cost is larger prediction sets. How large depends on what the laboratories share to choose these values, and we prove three results about it. First, without shared information, identical prediction sets are impossible. Second, a shared random seed makes them possible, but at a cost that no seed-based threshold rule can avoid in the worst case. Third, a small public sample from the same population removes the need for a seed, lowers the cost, and gives guarantees that hold for every continuous score distribution. We implement our method on ImageNet, CIFAR-100 and four language models, GPT-2, Pythia, Qwen and Llama, and confirm the theory. On ImageNet, the chance of identical prediction sets rises from zero to above ninety percent, for prediction sets about a fifth larger. On CIFAR-100, with six times less calibration data, the same pattern holds at a higher cost. On the language models, a shared seed reaches the chosen probability of identical prediction sets in all 42 settings, and its cost falls as the calibration data grow. In a question-answering service built on Llama, two laboratories whose stored predictions differ on more than a third of the questions under standard calibration end up with identical stored predictions ninety-five percent of the time, at the cost of storing larger prediction sets.

1 INTRODUCTION

Split-conformal prediction calibrates a trained model on held-out data and returns, for each input, a set of labels that contains the true label with a guaranteed probability (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2023). Consider two laboratories that calibrate the same approved model on their own data before a joint deployment: each computes a cutoff from its data and reports every label scored below it, and both maps carry the coverage guarantee. Yet the cutoff is an order statistic of continuous scores, so two independent calibrations differ with probability one, and the laboratories report different sets for some inputs. A regulator who reruns the calibration in an audit obtains a third map, and a serving system cannot reuse at one site the answer sets cached at another.

We ask what two analyst (or laboratories above or different sites), each calibrating, each calibrating the same model on n points from a common population, must share so that they deploy the same prediction-set map with probability at least $1 - \rho$, and how much larger the sets must be under each kind of shared information: nothing, a seed, a public sample (a *pilot*: n_p labeled points from the analysts' population, held by neither analyst), or the data. *Local validity:* each analyst's marginal coverage must be at least $1 - \alpha$ by virtue of its own calibration data, for every value of what is shared. *No exchange of cutoffs:* the analysts agree on a public protocol before calibrating and then calibrate alone. Pooling the data, broadcasting one analyst's cutoff, and deploying a threshold computed from the public sample alone all give identical maps, and all three fail the first requirement: their coverage

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: **Shared information and the cost of identity.** Identity is $\mathbb{P}(\tau_A = \tau_B)$, for a pilot from the analysts' population in the pilot rows. The cap holds with probability at least $1 - \rho$, for $\rho \le 1/32$ ($r_\rho = \sqrt{\log(1/\rho)}\,(n^{-1/2} + n_p^{-1/2})$, κ a local density ratio, $D_n \approx 1.13\sigma_F$ the mean distance between two Beta coverage levels, H_n, V_j and $\Gamma_{n,n_p}(b)$ as in Sections 2 to 5). The last three rows fail local validity; the last column says whose population the guarantee refers to.

Shared	Rule	Identity	Over-coverage cap	Valid for
Nothing	any two-sided ε-valid rule (Thm. 1)	$\le 1 - \min(\alpha, 1-\alpha) + \varepsilon$	$\le \varepsilon$ (assumed)	
Seed	score-unit rounding, margin known (Cor. 3)	$\ge 1 - \rho$	$O(\kappa\sigma_F/\rho)$	every seed
Seed and pilot	coverage-unit rounding (Thm. 2, Rem. 2)	$\ge 1 - \rho$	$(D_n + n_p^{-1})/\rho + O(r_\rho)$	every pilot
Pilot	pilot floor (Prop. 1)	$= \mathbb{E}[H_n(V_j)^2]$	$O(r_\rho)$	every pilot
Pilot	centred grid (Thm. 3)	$\ge \Gamma_{n,n_p}(b)$	$O(r_\rho)$	every pilot
Pilot	pilot-only threshold	1	$O(\sqrt{\log(1/\rho)/n_p})$	representative pilot
One cutoff	broadcast one threshold	1	$O(\sqrt{\log(1/\rho)/n})$	sender's population
Data	pooled calibration	1	$O(\sqrt{\log(1/\rho)/n})$	pooled population

rests on data that the analyst did not draw. Table 1 summarizes what each kind buys under both requirements.

In this paper we introduce calibration rules that round each analyst's cutoff upward onto a shared grid (Algorithm 1, Figure 1). Coverage can only grow when the cutoff grows, so upward rounding keeps local validity for every grid, and two analysts deploy the same map when their cutoffs fall in the same cell. The rules differ in how the grid is placed, which is what the analysts share. The analysis is distribution-free because the grid is laid out in units of coverage: each cutoff's coverage has the Beta law of an order statistic for every continuous score distribution, so the certified probability that two cutoffs share a cell depends on (n, n_p, α) and the grid alone.

- *Nothing shared, or a seed.* No rule that shares nothing and whose coverage is pinned within ε of the target for every distribution has worst-case mismatch below $\min(\alpha, 1-\alpha) - \varepsilon$ (Theorem 1). A shared seed and a map built from a representative pilot suffice: in coverage units the mismatch probability is at most the mean distance between two Beta coverage levels plus twice the map's distance to the pilot's empirical distribution function, divided by the grid width, with no assumption on the score density (Theorem 2).
- *The seed's price.* Every ρ-replicable threshold rule that shares only a seed and is valid with high probability has, in the worst case, a high-probability over-coverage cap of at least the smaller of a constant and an amount of order $\sqrt{\alpha(1-\alpha)/n}/\rho$ (Theorem 4, Corollary 1), and a variant of seeded rounding set from the hard family's margin constants attains this order up to a density factor of at most $2/\min(\alpha, 1-\alpha)$.
- *A public sample replaces the seed.* With n_p public scores from the analysts' population, the pilot floor (the larger of the own cutoff and a fixed pilot order statistic) and the centred grid (a grid centred at the target level) have identity certificates that are the same numbers for every continuous score distribution, and over-cover by an amount of order $\sqrt{\log(1/\rho)}\,(n^{-1/2} + n_p^{-1/2})$ (Proposition 1, Theorem 3, Corollary 2).
- *Measurements.* On ImageNet and CIFAR-100 the three certified rules meet their certificates within 2.3 clustered standard errors at every ρ from .5 to .01, and on ImageNet with APS at $\rho = .01$ the seedless rules reach identity 1.000 in all 1,000 pairs with 21% to 24% larger sets, where seeded rounding deploys the full label set for most seeds. A public sample from another population shows which rule loses what: the pilot-only threshold loses coverage, the pilot floor loses identity, and the centred grid keeps both (Section 7).

For practice: analysts with a public sample from their population (Table S5) use the pilot floor, which costs the least, or the centred grid, which kept identity of at least .73 under the pilot shifts we tested while the floor's fell to 0; analysts without a trusted public sample share a seed, pay an over-coverage of order $1/\rho$, and certify identity only under known density bounds (Corollary 3, Proposition 2).

Related work. Conformal prediction supplies the coverage guarantee and the exact Beta law of the realized coverage (Vovk, 2012), training-conditional variants control coverage with high probability (Vovk, 2012; Bian & Barber, 2023; Park et al., 2019), and discretized conformal prediction (Chen et al., 2018) rounds the response to make full conformal computable, with no agreement guarantee. Learning theory formalizes replicability as returning the same output on two independent samples with shared randomness and achieves it by rounding onto a randomly shifted grid (Impagliazzo et al.,

2022), without asking what rounding costs in coverage or set size. The replicable-coin lower bounds (Impagliazzo et al., 2022; Hopkins et al., 2024) are the engine of our Theorem 4, which embeds a coin in a calibration problem with explicit constants and the binomial variance $\alpha(1-\alpha)$. The impossibility of list size one for coin-bias estimation (Dixon et al., 2023, Theorem 4.7) and the list-replicability theory of Chase et al. (2023; 2024) concern learning without shared randomness, and Theorem 1 is their calibration analogue. The replicable quantile of Kalavasis et al. (2024) is a binary search over a finite grid without exact marginal validity, and the semi-replicability of Hopkins et al. (2025) shares unlabeled samples between the two runs, whereas our pilot is labeled and its certificates hold for every continuous score distribution. Other replicable algorithms share only a seed (Bun et al., 2023; Liu & Ye, 2024; Karbasi et al., 2023; Aamand et al., 2026; Eaton et al., 2026). Federated conformal prediction (Lu et al., 2023; Humbert et al., 2023; Plassier et al., 2023) communicates after seeing the data, which our second requirement excludes; conformal prediction under shift (Tibshirani et al., 2019; Podkopaev & Ramdas, 2021) corrects one analyst's coverage, not agreement. Conformal stability controls small changes in one map (Ndiaye, 2022; Bashari et al., 2023; Scholten & Günnemann, 2025; Min et al., 2026) and leaves open the equality of two independent calibrations. Full conformal prediction (Proposition 8), differing scores (Corollary 5), all proofs and Tables S1 to S22 are in the appendix.

2 PRELIMINARIES

Split conformal. $(X_i, Y_i) \overset{\text{iid}}{\sim} P$ on $\mathcal{X} \times \mathcal{Y}$. The score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is fixed and measurable, $S_i = s(X_i, Y_i)$ has a continuous distribution function F (Remark 1, Appendix C), and $q = F^{-1}(1-\alpha)$. With $k = \lceil (1-\alpha)(n+1) \rceil \le n$, the split-conformal cutoff is $\widetilde{\tau}_n = S_{(k)}$ ($\widetilde{\tau}$ below), the deployed prediction-set map is $\mathcal{C}_\tau(x) = \{y : s(x,y) \le \tau\}$, and $\mathbb{P}(Y_{n+1} \in \mathcal{C}_{\widetilde{\tau}_n}(X_{n+1})) = k/(n+1) \ge 1-\alpha$. The realized coverage $F(\widetilde{\tau}_n)$ has exactly the $\mathrm{Beta}(k, n{+}1{-}k)$ law (Vovk, 2012). We write H_n for its distribution function, $\sigma_F := \sqrt{\alpha(1-\alpha)/n}$ for its standard deviation to first order, the same for every F, and $D_n := \mathbb{E}|U_A - U_B|$ for U_A, U_B i.i.d. $\mathrm{Beta}(k, n{+}1{-}k)$. Then $D_n = (2/\sqrt{\pi})\,\sigma_F(1+o(1))$ and $D_n \le B_n^+ := \sqrt{2\alpha(1-\alpha)/n} + \sqrt{2}/n$. The *over-coverage*, or coverage inflation, of a deployed cutoff τ is $F(\tau) - (1-\alpha)$; a bound on it that holds with high probability is a *cap*.

Replicability. A calibrator $\mathcal{A}$ maps $(\mathcal{D}, r)$ to an output, with r a random string shared by the analysts. It is *ρ-replicable* over a class $\mathcal{P}$ if $\mathbb{P}_{\mathcal{D},\mathcal{D}' \sim P^{\otimes n}, r}[\mathcal{A}(\mathcal{D}; r) \ne \mathcal{A}(\mathcal{D}'; r)] \le \rho$ for all $P \in \mathcal{P}$ (Impagliazzo et al., 2022) (when the runs share nothing, each draws its own r), and *(ℓ, δ)-list replicable* if for every P some list $\mathcal{L}(P)$ with $|\mathcal{L}(P)| \le \ell$ has $\mathbb{P}[\mathcal{A}(\mathcal{D}; r) \in \mathcal{L}(P)] \ge 1 - \delta$ (Dixon et al., 2023). A calibrator is *seedless* if the analysts share no random string whose law is free of P: any randomness it uses is drawn by each analyst alone, and the only shared input, if any, is a public sample from P (Section 5). Theorem 1 concerns calibrators that share nothing. *Local validity* of a rule means that each analyst's marginal coverage, conditionally on everything shared, is at least $1 - \alpha$.

Definition 1 (Replication spectrum)**.** *Let two runs produce random set maps $\mathcal{C}, \mathcal{C}'$ and let X be an independent test input. We report* map identity $I := \mathbb{P}[\mathcal{C} \equiv \mathcal{C}']$ *and* pointwise agreement $A_{\mathrm{pt}} := \mathbb{P}[\mathcal{C}(X) = \mathcal{C}'(X)]$*, where probabilities include both calibration samples, the shared randomness and X (Milani Fard et al., 2016; Marx et al., 2020; Prastalo & Jablonka, 2026). For threshold rules,* threshold identity $I_\tau := \mathbb{P}[\tau = \tau']$ *satisfies $I_\tau \le I \le A_{\mathrm{pt}}$, and $1 - I_\tau$ is the* mismatch *probability. Below, identity means I_τ.*

Identity is the target for audits and caches, which need the same map, and targeting it can raise the expected number of labels on which two sets differ, since rounding concentrates disagreement into rare events that move a whole cell (Table S21).

3 IMPOSSIBILITY WITHOUT SHARED INFORMATION

Let $\mathcal{Y}$ be finite, $|\mathcal{Y}| = K \ge 2$, and $\mathcal{P}_{\mathrm{ac}}$ the class of distributions on $\mathbb{R}^d \times \mathcal{Y}$ with absolutely continuous X-marginal. A set-valued procedure is *uniformly two-sided ε-valid* if $\mathbb{P}_{\mathcal{D},r,(X,Y)\sim P}[Y \in \widehat{\mathcal{C}}(X)] \in [1-\alpha-\varepsilon,\, 1-\alpha+\varepsilon]$ for every $P \in \mathcal{P}_{\mathrm{ac}}$. Two-sidedness excludes the trivial rule $\widehat{\mathcal{C}} \equiv \mathcal{Y}$.

Theorem 1 (No replication without shared information)**.** *If $\varepsilon + \delta < \min(\alpha, 1-\alpha)$, no $(1, \delta)$-list-replicable procedure is uniformly two-sided ε-valid. Every ρ-replicable procedure whose two*

runs share nothing (no seed and no public sample) is $(1, \rho)$-list replicable, so no such procedure is both uniformly two-sided ε-valid and ρ-replicable over $\mathcal{P}_{\mathrm{ac}}$ when $\varepsilon + \rho < \min(\alpha, 1 - \alpha)$. (Proof: Appendix A.)

A public pilot is distribution-dependent information, which Section 5 shows makes seedless replication possible, and any calibrator whose output law is atomless given the shared randomness, such as a bagged, fold-aggregated or kernel-smoothed cutoff, has identity zero (Lemma 1).

4 Rounding in coverage units with a shared seed

Algorithm 1 states the rule and Figure 1 shows the two grids and the anchor. Its upper inverse Q_G^+ never moves a cutoff down, even across flat stretches of G, so coverage is preserved. The rules of this paper differ only in what is shared. *Seeded rounding* shares a random offset and a width, in *score units* (G the identity, RECAL in the appendix) or in *coverage units* (G a public map); the *centred grid* shares a pilot and centres a cell at $1 - \alpha$; and the *pilot floor* shares a pilot and a single anchor (both in Section 5). In practice a seeded rule needs the pilot too, to set its width, so every rule we run shares a pilot; the seed-only regime appears only in the lower bound of Section 6.

Theorem 2 (Rounding in the units of a public map)**.** *Run Algorithm 1 with a seeded phase independent of all data.* (i) Validity for every map. *For every realization of G and u, each analyst's deployed cutoff satisfies $\tau \geq \widetilde{\tau}$, so its marginal coverage is at least $k/(n+1) \geq 1 - \alpha$ conditionally on (G, u), and equal rounded levels give identical deployed maps.* (ii) Mismatch. *Conditionally on G and the raw cutoffs, $\mathbb{P}_u(\tau_A \neq \tau_B) \leq \min\{|G(\widetilde{\tau}_A) - G(\widetilde{\tau}_B)|/b, 1\}$, with equality for the rounded levels.* (iii) Two information regimes. *(a) If $G = F$, then $\mathbb{P}(\tau_A \neq \tau_B) \leq D_n/b$. (b) If G is built from a pilot of n_p i.i.d. draws from F, independent of both calibration samples, with $\|G - \widehat{F}_{n_p}\|_\infty \leq \lambda/n_p$ for the pilot's empirical distribution function $\widehat{F}_{n_p}$, then, averaging over the pilot, $\mathbb{P}(\tau_A \neq \tau_B) \leq (D_n + 2\lambda/n_p)/b$.* (iv) Coverage cost. *If $G = F$, then each analyst has $0 \leq F(\tau) - F(\widetilde{\tau}) \leq b$ and $\mathbb{E}_u[F(\tau) - F(\widetilde{\tau}) \mid \widetilde{\tau}] \leq b/2$. (Proof: Appendix C.)*

Why coverage units remove the density. Rounding in score units replaces D_n by $\mathbb{E}|\widetilde{\tau}_A - \widetilde{\tau}_B|$, approximately D_n divided by the local density, which needs a density lower bound to control. In coverage units both cutoffs' levels $F(\widetilde{\tau})$ are $\mathrm{Beta}(k, n{+}1{-}k)$ variables whatever F is, so their mean distance is the number D_n, the width D_n/ρ gives mismatch at most ρ against the oracle map, and neither case of (iii) uses a density, margin or density-ratio assumption. With the pilot map of the experiments, the pilot's empirical distribution function interpolated through its midpoints ($\lambda = \frac{1}{2}$, Appendix F), the width $(D_n + 1/n_p)/\rho$ carries certified mismatch ρ. In score units the same rule needs a density: a margin condition with density ratio $\kappa := f_{\max}/f_{\min}$ near the quantile gives a sufficient sample size of leading order $32\kappa^2\alpha(1-\alpha)/(\varepsilon\rho)^2$ for coverage accuracy ε (Theorem 5, Corollary 3), and the *plug-in width* $\sqrt{2\alpha(1-\alpha)/n}/(\hat{f}\rho)$, with $\hat{f}$ a pilot estimate of the density at the quantile, is free of κ asymptotically (Proposition 2) but carries no finite-sample certificate, because κ cannot be certified from data. It is the rule of the language-model and cache runs in Section 7.

5 Seedless calibration with a public pilot

Write $Z_{(1)} \leq \cdots \leq Z_{(n_p)}$ for the pilot's order statistics. Deploying one of them as the threshold gives identity one and coverage that rests entirely on the pilot. The *pilot floor* adds local-validity insurance:

Algorithm 1 Rounding onto a shared grid in the units of a public map

Require: Shared by all analysts: the frozen score s (with the deterministic jitter of Remark 1 when scores have atoms), the level α, a public right-continuous nondecreasing map G fixed independently of all calibration and test data (a pilot's distribution function, the true F as an oracle benchmark, or the identity for score units), a width $b > 0$ in the units of G (β in the appendix for score units), and a phase u: *seeded*, $u \sim \mathrm{Unif}[0, b)$ from a shared seed, or *seedless*, $u_c := (1 - \alpha + b/2) \bmod b$, which centres a cell at $1 - \alpha$.

1: Each analyst, on its own n calibration scores: $\widetilde{\tau} \leftarrow S_{(k)}$ with $k = \lceil (1-\alpha)(n+1) \rceil$.
2: $v \leftarrow \mathrm{rd}_{u,b}(G(\widetilde{\tau}))$, where $\mathrm{rd}_{u,b}(x) := u + b\lceil (x-u)/b \rceil$ is the smallest grid point $\geq x$.
3: $\tau \leftarrow Q_G^+(v) := \inf\{t : G(t) > v\}$, with $\tau = +\infty$ (the full label set) if no such t exists.
4: Deploy $\mathcal{C}_\tau(x) = \{y : s(x, y) \leq \tau\}$.

each analyst deploys the pilot's order statistic unless its own cutoff is higher, and at $\rho = .1$ in Section 7 it uses almost the same rank as the pilot-only threshold (1,822 against its training-conditional rank 1,823 at $\delta = .05$, of 2,000), so it costs the same.

Proposition 1 (Pilot floor). *Fix a rank $j \le n_p$, let $T := Z_{(j)}$, and let each analyst deploy $\tau := \max\{\widetilde{\tau}, T\}$.* (i) *For every pilot law and every pilot realization independent of the analyst's data, the analyst's marginal coverage is at least $1-\alpha$.* (ii) *If the pilot is i.i.d. from the analysts' continuous law F and independent of both calibration samples, then $\mathbb{P}(\tau_A = \tau_B) = \mathbb{E}[H_n(V_j)^2]$ exactly, where $V_j \sim \mathrm{Beta}(j, n_p{+}1{-}j)$, and $\mathbb{E}F(\tau) = \mathbb{E}\max\{U, V_j\}$ with $U \sim \mathrm{Beta}(k, n{+}1{-}k)$ independent of V_j.* (iii) *Under the assumptions of (ii), with $j := \lceil n_p(k/n + t_n + \bar{t}_{n_p})\rceil$, $t_n := \sqrt{\log(8/\rho)/(2n)}$ and $\bar{t}_{n_p} := \sqrt{\log(4/\rho)/(2n_p)}$, if $j \le n_p$ then both analysts deploy T and $1-\alpha \le F(T) \le 1-\alpha+t_n+2\bar{t}_{n_p}+2/n+1/n_p$ with probability at least $1-\rho$.* (iv) *For a pilot from any law, conditionally on T the threshold identity equals $H_n(F(T))^2$, which is at most ρ^2 once $F(T) \le k/(n+1) - \sqrt{\log(1/\rho)/(2n)}$. (Proof: Appendix C.)*

The identity in (ii) depends only on (n, n_p, α, j), so the smallest rank with identity at least $1-\rho$ is computable before any data are seen: at $(n, n_p) = (15{,}000, 2{,}000)$ it is $j = 1{,}822$, $1{,}826$ and $1{,}835$ for $\rho = .1, .05, .01$, with exact mean coverage .911, .913 and .917 (cf. Table 2). The value in (ii) averages over the pilot: given the pilot, identity is $H_n(F(T))^2$ by (iv), near 0 or 1 unless the anchor lies within a few σ_F of $1-\alpha$ (Appendix F). Part (iv) is the rule's weakness (identity vanishes once $F(T)$ lies a few σ_F below $1-\alpha$, while coverage stays valid by (i)), and the centred grid rounds every cutoff up onto a grid centred at the target level, so cutoffs above the anchor also coincide.

Theorem 3 (Centred grid with a pilot: a distribution-free certificate). *Run Algorithm 1 with the seedless phase u_c, and let G, with values in $[0,1]$, be built from a pilot of n_p scores, independent of both calibration samples, with $\|G - \widehat{F}_{n_p}\|_\infty \le \lambda/n_p$ for some $\lambda \ge 0$. Put $i_a := \lfloor n_p(1-\alpha-b/2)+\lambda\rfloor + 1$ and $i_c := \lfloor n_p(1-\alpha+b/2)-\lambda\rfloor + 1$, and let V_a and V_c be the i_a-th and i_c-th order statistics of the same n_p i.i.d. $\mathrm{Unif}(0,1)$ variables, with the conventions $V_{(i)} := 0$ for $i \le 0$ and $V_{(i)} := 1$ for $i \ge n_p + 1$.* (i) *For every pilot law and every pilot realization, each analyst's marginal coverage is at least $1-\alpha$, provided the pilot is independent of that analyst's calibration and test points.* (ii) *If the pilot is i.i.d. from the analysts' continuous law F, then $\mathbb{P}(\tau_A = \tau_B) \ge \Gamma_{n,n_p}(b) := \mathbb{E}\big[(H_n(V_c) - H_n(V_a))_+^2\big]$. The certificate $\Gamma_{n,n_p}(b)$ depends only on $(n, n_p, \alpha, b, \lambda)$, and so holds for every continuous F.* (iii) *If, in addition, $b/2 \ge t_n + \bar{t}_{n_p} + (\lambda+1)/n_p + 2/n$ with $t_n, \bar{t}_{n_p}$ as in Proposition 1, then with probability at least $1-\rho$ both analysts deploy one common threshold τ, and $1-\alpha \le F(\tau) \le 1-\alpha+b$.* (iv) *For a pilot from any law, conditionally on the pilot, the two analysts' rounded levels coincide with probability $\sum_m \pi_m(G)^2$, where $\pi_m(G)$ is the probability that $G(\widetilde{\tau})$ falls in the m-th grid cell (cells at levels ≥ 1 merged). This lower-bounds the probability of identical maps, with equality of thresholds when G is continuous and strictly increasing. (Proof: Appendix C.)*

Why centring makes the seed unnecessary. Both transformed cutoffs concentrate within a few σ_F of $1-\alpha$, so a cell of width $b \gg \sigma_F$ centred there catches both with high probability, and the certificate in (ii) is a lower bound on the probability that both fall in that central cell $(1-\alpha-b/2, 1-\alpha+b/2]$, where both deploy $Q_G^+(1-\alpha+b/2)$, so the measured identity can exceed it. At $(n, n_p) = (15{,}000, 2{,}000)$ the smallest certified widths for $\rho = .1, .05, .01$ are $b = .0265, .0306, .0396$, against $.033, .065, .326$

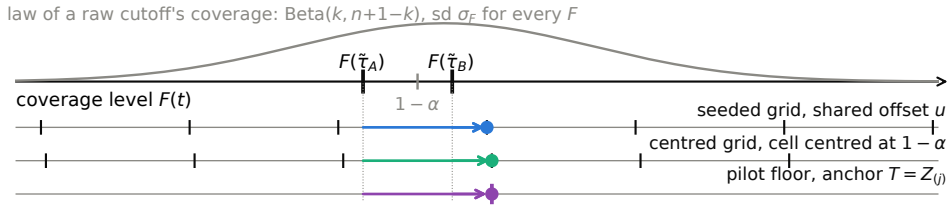


Figure 1: The mechanism (schematic). A public map G sends scores to coverage levels, where each raw cutoff's level $F(\widetilde{\tau})$ has the Beta law with standard deviation σ_F whatever F is (under a pilot map the axis carries $G(\widetilde{\tau})$). A grid of width b is laid on this axis and each analyst rounds its level up to the next grid point, so both deploy the same threshold when their levels share a cell: the seeded grid has a shared random offset u, the centred grid a cell centred at $1-\alpha$, and the pilot floor a single anchor T, a pilot order statistic, which both deploy when their cutoffs lie below it.

for seeded rounding. Part (iv) describes a shifted pilot: coverage holds by (i), and identity follows a lattice law with the pilot's *level gap* $G(q) - (1-\alpha)$ as the seed phase (Proposition 4, Section 7).

6 Lower bounds in the seed-only regime

Let $\mathcal{P}_2(\alpha, \varepsilon)$ be the two-piece family of laws P_m on $[0,1]$ with density $2m$ on $[0, \frac{1}{2})$ and $2(1-m)$ on $[\frac{1}{2}, 1]$, for $|m - (1-\alpha)| \le 2\varepsilon$, with $\varepsilon \le \min(\alpha, 1-\alpha)/4$ (Appendix D). Every P_m is atomless and satisfies the margin condition of Appendix B with $f_{\min} \ge \min(\alpha, 1-\alpha)$, $f_{\max} \le 2$ and $\Delta = \frac{1}{3}$, so its density ratio is at most $2/\min(\alpha, 1-\alpha)$.

Theorem 4 (Sample-size lower bound)**.** *Fix an integer $n \ge 1$ and $\alpha \in (0,1)$, $0 < \varepsilon \le \min(\alpha, 1-\alpha)/4$, $0 < \rho \le 1$, $0 \le \delta < \frac{1}{4}$. Let $\mathcal{A} : [0,1]^n \times \mathsf{R} \to [-\infty, +\infty]$ be jointly measurable, with $F_P(-\infty) := 0$ and $F_P(+\infty) := 1$, with a shared seed r drawn from a law ν on a measurable space R, where ν does not depend on the data distribution, and let $\mathcal{P}$ be any class of laws on $[0,1]$ that contains $\mathcal{P}_2(\alpha, \varepsilon)$. If $\mathcal{A}$ is ρ-replicable over $\mathcal{P}$ and $\mathbb{P}[F_P(\mathcal{A}(\mathcal{D}; r)) \in [1-\alpha-\varepsilon, 1-\alpha+\varepsilon]] \ge 1-\delta$ for every $P \in \mathcal{P}$ with distribution function F_P, then $n \ge \frac{1}{72}\,\alpha(1-\alpha)/(\varepsilon^2\rho^2)$ for $\delta \le \frac{1}{32}$; the general constant, which tends to $\pi^2/128$ as $\delta \to 0$, is in Appendix D. (Proof: Appendix D.)*

Corollary 1 (Coverage-inflation lower bound)**.** *In the setting of Theorem 4, let $0 \le \delta \le \frac{1}{32}$, $0 < e \le \min(\alpha, 1-\alpha)/4$, and let $\mathcal{P}$ contain $\mathcal{P}_2(\alpha - e/2, e/2)$. If $\mathcal{A}$ is ρ-replicable over $\mathcal{P}$ and, for every $P \in \mathcal{P}$, satisfies $\mathbb{P}[F_P(\mathcal{A}(\mathcal{D}; r)) \ge 1-\alpha] \ge 1-\delta$ and $\mathbb{P}[F_P(\mathcal{A}(\mathcal{D}; r)) \le 1-\alpha+e] \ge 1-\delta$, then $e \ge \frac{3}{20}\sqrt{\alpha(1-\alpha)/n}\,/\rho$ (the sharper constant, in $\alpha' := \alpha - e/2$, is in Appendix D). (Proof: Appendix D.)*

The corollary bounds a worst-case high-probability cap on coverage inflation for caps below $\min(\alpha, 1-\alpha)/4$, and the order is attained: the variant RECAL$^+$ of seeded score-unit rounding, which raises the target level by a DKW margin and sets its width from the margin constants of $\mathcal{P}_2$, has cap $O(\sqrt{\log(1/\delta)/n} + \kappa\sigma_F/\rho)$ with $\kappa \le 2/\min(\alpha, 1-\alpha)$ (Proposition 3). The cap transfers to mean set size, since the excess mean cardinality is at least the excess coverage, and to deployed maps whenever distinct admissible thresholds give different maps (Lemma 4).

Corollary 2 (What a representative pilot changes)**.** *Let $\rho \le 1/32$. With the width of Theorem 3(iii) or the rank of Proposition 1(iii) (when that rank is at most n_p), a seedless rule with a representative pilot is ρ-replicable for every continuous F, satisfies $\mathbb{P}\{F(\tau) \ge 1-\alpha\} \ge 1-\rho$, and has an inflation cap $e_{\text{pilot}} = O\big(\sqrt{\log(1/\rho)}\,(n^{-1/2} + n_p^{-1/2})\big)$ with the same probability. Any threshold calibrator with shared randomness independent of F that meets these requirements for every F in a class containing $\mathcal{P}_2(\alpha - e/2, e/2)$ has $e \ge \min\{\frac{3}{20}\sqrt{\alpha(1-\alpha)/n}/\rho,\ \min(\alpha, 1-\alpha)/4\}$ (Corollary 1 with $\delta = \rho$), and $e \ge \min\{\frac{3}{20}\sqrt{\alpha(1-\alpha)/(n+n_p)}/\rho,\ \min(\alpha, 1-\alpha)/4\}$ when its analysts each draw $n + n_p$ samples. When the first term is smaller and $n_p \le n$, $e_{\text{pilot}}/e = O\big(\rho\sqrt{\log(1/\rho)}\,(1 + \sqrt{n/n_p})\big)$ for both comparators. (Proof: Appendix D.)*

Corollary 2 is a rate statement in ρ, and its constants do not bite at the parameters of Section 7: at $(n, n_p, \rho) = (15{,}000, 2{,}000, .01)$ the seed-only lower bound on the cap is .025, below the certified seedless width .040. The separation Section 7 measures comes instead from Theorem 2(ii), an equality for the rounded levels: seeded rounding needs the width $(D_n + 1/n_p)/\rho = .33$ for mismatch .01, eight times the certified seedless width .040.

7 Experiments

Theorems 1 and 4 are worst-case statements that no measurement can test; the experiments test the rules of Sections 4 and 5. They test that the certificates of Theorems 2 and 3 and Proposition 1 hold at their widths and ranks, that each rule's coverage inflation follows the predicted curve as ρ decreases, with the seed's $1/\rho$ against the pilot's $\sqrt{\log(1/\rho)}$, that a pilot from a shifted population affects coverage and identity as Theorem 3(i),(iv) and Proposition 1(i),(iv) predict, and that the seed's price behaves as predicted on language-model vocabularies. The *premium* of a rule is the relative increase of its mean set size over standard split conformal with the same score. The four rules of Section 4, at the widths and ranks stated in the caption of Table 2, run on the ImageNet validation logits of a frozen ResNet-50 (Northcutt et al., 2021) (N=50,000) and on a synthetic pool at ImageNet dimensions,

Table 2: **Certificates against measurements on real ImageNet** (master protocol, 1,000 pairs per setting, ± one standard error clustered by pilot redraw). Rules are at the width or rank for the row's ρ (plug-in width for score units, $(D_n + 1/n_p)/\rho$ for seeded coverage units, smallest certified width for the grid, certified rank for the floor). Certified: the lower bound of Theorem 2(iii-b) or 3(ii), or the exact value of Proposition 1(ii). Coverage: held-out. Premiums are relative to standard split conformal on the same pairs (identity .000, APS coverage .9018). Fresh i.i.d.: the range of measured identity over five continuous laws, 5,000 independent pilot and calibration draws each (Table S7). $1.000 \pm .000$: every pair of every redraw matched (20 redraws without a mismatch bound the per-redraw mismatch probability by .14 at 95%). The score-unit rule has no certificate and is included because it is the rule of Table 3 and of Appendix F; premiums in the thousands of percent mean that most seeds deploy the full label set (+2,792% for APS).

			APS			RAPS		Fresh i.i.d.
Rule	ρ	Certified	Identity	Coverage	Premium	Identity	Premium	Identity
Pilot-only threshold		1	1.000	.9117	+12%	1.000	+5%	1.000
Seeded, score units	.1		$.925 \pm .012$	.9176	$+22\% \pm 0.7$	$.934 \pm .010$	$+13\% \pm 0.8$	
	.05		$.962 \pm .006$	.9328	$+55\% \pm 2$	$.967 \pm .005$	$+33\% \pm 2$	
	.01		$.992 \pm .001$	.9837	$+1{,}830\% \pm 15$	$.993 \pm .001$	$+166\% \pm 8$	
Seeded, cov. units	.1	$\geq .900$	$.906 \pm .014$	.9174	$+22\% \pm 1.0$	$.911 \pm .014$	$+9\% \pm 0.4$	.914–.916
	.05	$\geq .950$	$.953 \pm .007$	.9335	$+58\% \pm 2$	$.955 \pm .007$	$+61\% \pm 7$	.957–.958
	.01	$\geq .990$	$.991 \pm .001$	.9850	$+1{,}992\% \pm 10$	$.991 \pm .001$	$+12{,}752\% \pm 89$	.991–.992
Centred grid	.1	$\geq .902$	$.977 \pm .017$	.9137	$+15\% \pm 2$	$.918 \pm .055$	$+6\% \pm 0.9$	.939–.950
	.05	$\geq .952$	$.999 \pm .001$	.9155	$+18\% \pm 2$	$.949 \pm .049$	$+8\% \pm 1$	.966–.974
	.01	$\geq .991$	$1.000 \pm .000$	.9194	$+24\% \pm 2$	$1.000 \pm .000$	$+10\% \pm 0.9$	.992–.994
Pilot floor	.1	$= .912$	$.949 \pm .048$	.9113	$+12\% \pm 1$	$.855 \pm .073$	$+5\% \pm 0.8$	.904–.914
	.05	$= .951$	$.964 \pm .035$	.9132	$+14\% \pm 2$	$.947 \pm .050$	$+6\% \pm 0.8$	.946–.952
	.01	$= .990$	$1.000 \pm .000$	.9173	$+21\% \pm 2$	$.961 \pm .039$	$+8\% \pm 0.9$	.990–.991

with the adaptive (APS) and regularized adaptive (RAPS) scores (Romano et al., 2020; Angelopoulos et al., 2020) (APS only on the synthetic pool), next to the pilot-only threshold and the baselines of Table S9. The same protocol runs on CIFAR-100 (Northcutt et al., 2021) at $(n, n_p) = (2{,}460, 937)$; the uncertified score-unit rule alone runs on fine-tuned CIFAR-100 models (Table S10), next-token sets of four language models on WikiText-103 (Table 3), an answer cache on WebQuestions (Quach et al., 2024) (Table S17), five states of the Folktables income task (Tables S14 to S16), four hospitals and two small sets, digits and abalone (Appendix F).

Master protocol. On ImageNet we fix a 10,000-point held-out test split, draw a public pilot of n_p=2,000 from the rest, split the remainder into two disjoint analyst halves, and let each analyst draw n=15,000 points from its half without replacement: 1,000 analyst pairs nested in 20 pilot redraws per setting, standard errors clustered by redraw, every rule on the same pairs; other pools use the same design at their own sizes (Appendix F). On the image pools, seeded rules' identity, coverage and set size are integrated exactly over the shared offset (elsewhere over fixed offsets, Appendix F), seedless identity is the observed equality of the deployed thresholds, and coverage and size are measured on the held-out split: coverage is that of the raw set $\mathcal{C}_\tau$ of Algorithm 1, and size counts an empty raw set as one label, since a system deploys at least one. Every seeded rule takes its width from the pilot, so the master protocol instantiates the seed-and-pilot regime and does not test the seed-only regime of Theorem 4; the in-pool frontier of Appendix F probes only its rate.

Certificates and predictions (Table 2, Figure 2). At every ρ from .5 to .01 and on all five pool settings (Tables S2 to S4), the measured identity of seeded coverage-unit rounding is at or above $1 - \rho$, the centred grid lies above its certificate or within .5 clustered standard errors of it, and the pilot floor lies within 2.3 clustered standard errors of its exact value on either side, or above it where no pair mismatched, (below it in 24 of the 35 settings, a tendency inside the pilot-to-pilot variability of 20 redraws that a Gaussian model of the finite pool does not reproduce, Appendix F). Fresh i.i.d. draws remove the finite pool: the floor is then within 2.1 standard errors of its exact value at every ρ and every law, on both sides; the seeded rule's identity exceeds $1 - \rho$ by the slack $1/(n_p b)$ of the pilot term in Theorem 2(iii-b) to within .002; and the certificates hold at the nine (n, n_p) pairs of Table S8, at $\alpha \in \{.05, .1, .2\}$, and for five and ten analysts. Coverage inflation follows the predicted curves (Figure 2(c)): $b/2$ for the centred grid and $\mathbb{E}\max\{U, V_j\} - \mathbb{E}U$ for the floor, both of order $\sqrt{\log(1/\rho)}$, against $\mathbb{E}\min\{U + W, 1\} - \mathbb{E}U$ for seeded coverage-unit rounding, with W the rounding excess, uniform on $[0, b)$, which saturates at the full label set, and $\sqrt{2\alpha(1-\alpha)/n}/(2\rho)$ for the plug-in score-unit width, which exceeds the full set's inflation α below $\rho = .02$. The held-out

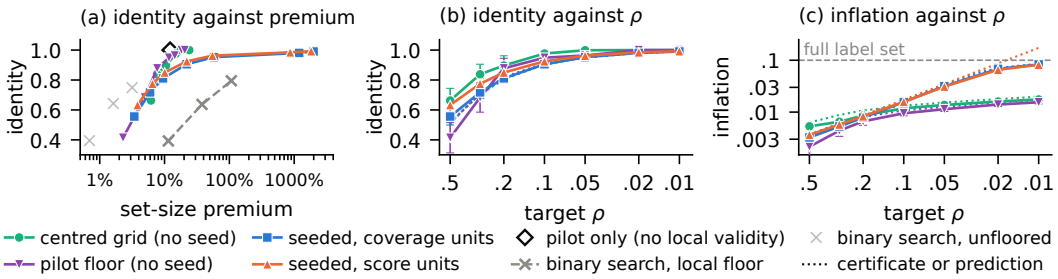


Figure 2: Theory against measurement on real ImageNet (APS, master protocol, $\rho \in \{.5, .3, .2, .1, .05, .02, .01\}$, bars one clustered standard error). (a) Identity against set-size premium (ρ decreasing left to right), with the pilot-only threshold and the floored (exactly valid) and unfloored binary searches at $\varepsilon = .02, .05, .1$. (b) Identity against ρ with the certificates dotted: $1 - \rho$, $\Gamma_{n,n_p}(b)$ and the exact $\mathbb{E}[H_n(V_j)^2]$. (c) Coverage inflation over standard split conformal on the same pairs, against the i.i.d. predictions of Section 7 (dotted) and the full label set's inflation (dashed).

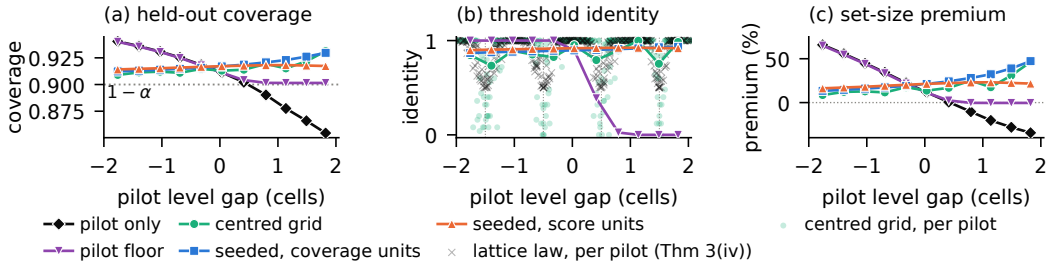


Figure 3: A public pilot from a shifted population (real ImageNet, APS, $\rho = .1$, analysts' data fixed, 40 pilots at each of eleven pilot compositions). Horizontal axis: the pilot's level gap $G(q) - (1 - \alpha)$ at the analysts' quantile, in cells of the centred grid (width .0265), positive when the pilot over-represents the small-score classes. (a) Held-out coverage, nominal level dotted. (b) Threshold identity (the pilot-only threshold, at 1, is not drawn), with the centred grid's per-pilot values and lattice predictions (Theorem 3(iv)), dotted lines at half-integer gaps. (c) Set-size premium.

coverage of the floor and the grid is within .005 of its prediction at every ρ on all five settings, inside the sampling error of the fixed test split, and within .0005 under fresh draws (Appendix F). The same appendix fits the score-unit rule's own mismatch frontier, in its width parameter ε, as $(\varepsilon\rho)^{-\gamma}$, with $\gamma = 2.09$ $[1.99, 2.28]$ on the synthetic pool and 1.83 $[1.79, 1.87]$ on the finite ImageNet pool; it is a description, not a test of Theorem 4.

The cost of identity. Table 2 and Figure 2(a) show the two orders of Sections 4 and 5 side by side. At $\rho = .1$ every certified rule costs between $+12\%$ and $+22\%$ with APS, and the pilot floor costs the same as the pilot-only threshold ($+12\%$) while adding local validity. As ρ falls the seeded rules pay the seed's price: $+55\%$ and $+58\%$ at $\rho = .05$, and at $\rho = .01$ their widths exceed α in coverage units (.33, and about .35 for the plug-in score-unit width), so most seeds deploy the full label set and the premiums approach the full set's $+2{,}792\%$ at $+1{,}830\%$ and $+1{,}992\%$. The seedless rules pay $\sqrt{\log(1/\rho)}$: the floor and the grid reach identity 1.000 in all 1,000 pairs at $\rho = .01$ (exact value .990 for the floor, certified $\geq .991$ for the grid) at $+21\%$ and $+24\%$. With RAPS the seedless premiums are at most $+10\%$ at every ρ, the grid reaches 1.000 at $\rho = .01$ and the floor $.961 \pm .039$, against $+166\%$ for seeded score-unit rounding at identity .993 (Table S3), with the same ordering on the synthetic pool (Table S2). On CIFAR-100, where $n = 2{,}460$ makes every width larger, the ordering is the same at larger premiums, with one exception, RAPS at $\rho \geq .2$, where the seeded score-unit rule has higher identity than the grid at a lower premium (Table S4). Replicable binary search (Kalavasis et al., 2024), with an upward-shifted target and floored at the local cutoff so that it is exactly valid, reaches identity .39, .64 and .80 at $+11\%$, $+38\%$ and $+107\%$, each point dominated by a certified rule; with a symmetric target and no floor it is cheaper (.40, .64 and .75 at $+1\%$ to $+3\%$) but undercovers in 32 to 46 percent of pairs (standard split conformal: 13 percent), and $n = 15{,}000$ is far below the sample size of its guarantee (Appendix F). Other baselines are in Tables S2 and S9.

Table 3: **Next-token prediction sets of four language models** (WikiText-103, master protocol, seeded score-unit rounding at the plug-in width, $\rho = .1$, $n = 2{,}625$, or $6{,}000$ for Llama-3.2-1B). Each cell: identity over $1{,}000$ pairs in 10 pilot redraws (standard errors over pairs at most .0018; per-redraw values were not stored, so errors are not clustered, and clustered errors in the same design on ImageNet are about .01), and premium over standard split conformal, whose identity is .000.

Score	GPT-2 (K=50k)	Pythia-1.4B (K=50k)	Qwen2.5-1.5B (K=152k)	Llama-3.2-1B (K=128k)
APS	.939 (+143%)	.920 (+165%)	.914 (+183%)	.926 (+70%)
RAPS	.942 (+128%)	.944 (+130%)	.951 (+137%)	.942 (+102%)

A pilot from a shifted population (Figure 3, Table S6). This is the case the local-validity requirement of Section 1 is for: the shared object is wrong, and only the rules that floor or round a local cutoff are guaranteed to survive it. As the pilot's level gap moves from -1.76 to $+1.82$ cells, the pilot-only threshold's coverage moves from .940 to .855, below nominal for every pilot once the gap exceeds one cell, and its premium from $+66\%$ to -34%, while every rule that floors or rounds a local cutoff keeps coverage above .901 at every gap, as parts (i) of Theorems 2 and 3 and Proposition 1 require. The pilot floor follows the pilot-only threshold while the anchor lies above the analysts' quantile, with identity 1 at gaps of $-.32$ cells and below and the same premium, and loses identity as the anchor approaches and then falls below the quantile: .90, .39, .03 and 0 at gaps of .03, .41, .79 and beyond, as Proposition 1(iv) predicts. The centred grid's identity falls to .73 and .75 at gaps of -1.40 and 1.49 cells and rises to .95, 1.00 and .99 at .03, 1.14 and 1.82 cells, which the lattice law of Theorem 3(iv) tracks across pilots ($R^2 = .67$ for positive gaps and .60 for negative gaps), at a premium between $+9\%$ and $+46\%$, and seeded identity stays between .87 and .93 at every gap, so the grid keeps identity of at least .75 where the floor's is 0.

The seed's price at scale: language models and an answer cache (Table 3). These runs use the seeded score-unit rule alone, at the plug-in width, on vocabularies of 50k to 152k labels, so they show the price of Section 4 where it is steepest (the certified rules were not run on these pools; their certificates at these sizes are in Appendix F). The rule has mean identity of at least $1 - \rho$ in all 42 master-protocol settings of model, score, n and $\rho \in \{.1, .05, .01\}$, with mismatch between $.49\rho$ and $.86\rho$ against the $.80\rho$ that Proposition 2 predicts for an exact density estimate (three of the five in-pool Pythia runs of Table S13 fall below $1 - \rho$, within 1.4 standard errors). The price falls with n and rises with $1/\rho$, as the width $\propto 1/(\rho\sqrt{n})$ predicts: at $n \le 6{,}000$ and $\rho = .1$ the premium is $+70\%$ to $+183\%$; at n=13,500, with larger pools and pilots, it falls to $+33\%$ to $+35\%$ with APS and $+49\%$ to $+70\%$ with RAPS (Table S12); and at $\rho = .05$ and $n \le 6{,}000$ the mean APS set holds 8 to 42 percent of the vocabulary while RAPS costs $+205\%$ to $+258\%$, the onset of the saturation of Figure 2(c). On an answer cache on WebQuestions with two Llama-3.2 models (in-pool, one partition into a pilot and two sites, $\alpha = .48$ and .41, Appendix F), the same rule at 1.5 times the plug-in width keeps the two sites' caches identical with probability .948 and .973 at $\rho = .1$, where standard calibration leaves 39% and 17% of cached answer sets different, at 1.9 and 1.8 times the cache size.

8 Conclusion

We asked what two analysts calibrating the same model on their own data must share to deploy the same prediction-set map. We proved that sharing nothing cannot work, that a seed costs an over-coverage of order $1/\rho$ that no seed-only threshold rule can lower in the worst case, and that a public sample removes the seed at a cost of order $\sqrt{\log(1/\rho)}$, with certificates that hold for every continuous score law. We validated the theory on ImageNet, CIFAR-100 and four language models: the certificates hold at their widths and ranks, with APS the seedless rules reach identity one with sets about a fifth larger, and under a shifted public sample every rule of this paper keeps its coverage. The limitations are that the lower bounds cover scalar thresholds with distribution-independent randomness and leave a lower bound with a public sample open, that the certificates assume one population, and that the language-model and cache experiments use the uncertified score-unit rule only.

Reproducibility statement The supplementary material contains the code of the real-model study and of the verification experiments with their calibration statistics, the per-seed result files, the job logs and the script that regenerates every real-model table and figure.

AI-use statement Language-model assistants were used, under the authors' direction and review, for drafting and editing text, for implementing and debugging the experiment code. The authors retain responsibility for the theorems, proofs, experiments and claims.

Ethics statement The work analyses publicly released pretrained models on public datasets and involves no human subjects or deployment. The clinical and census experiments use the de-identified UCI Heart Disease files and public-use ACS microdata under their licences, and no attempt is made to re-identify individuals.

REFERENCES

Anders Aamand, Maryam Aliakbarpour, Justin Y Chen, Shyam Narayanan, and Sandeep Silwal. On the structure of replicable hypothesis testers. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 5771–5823. SIAM, 2026.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

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

Alexander A Balinsky and Alexander David Balinsky. When can we reuse a calibration set for multiple conformal predictions? *arXiv preprint arXiv:2506.19689*, 2025.

Meshi Bashari, Amir Epstein, Yaniv Romano, and Matteo Sesia. Derandomized novelty detection with fdr control via conformal e-values. *Advances in Neural Information Processing Systems*, 36: 65585–65596, 2023.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1533–1544, 2013.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, pp. 802–837, 2013.

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

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 520–527, 2023.

Zachary Chase, Shay Moran, and Amir Yehudayoff. Replicability and stability in learning. *arXiv preprint arXiv:2304.03757*, 2023.

Zachary Chase, Bogdan Chornomaz, Shay Moran, and Amir Yehudayoff. Local borsuk-ulam, stability, and replicability. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 1769–1780, 2024.

Wenyu Chen, Kelli-Jean Chun, and Rina Foygel Barber. Discretized conformal prediction for efficient distribution-free inference. *Stat*, 7(1):e173, 2018.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Peter Dixon, Aduri Pavan, Jason Vander Woude, and N_V Vinodchandran. List and certificate complexities in replicable learning. *Advances in Neural Information Processing Systems*, 36: 30784–30806, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Eric Eaton, Marcel Hussing, Michael Kearns, Aaron Roth, Sikata Sengupta, and Jessica Sorrell. Replicable reinforcement learning with linear function approximation. In *International Conference on Learning Representations*, volume 2026, pp. 126582–126616, 2026.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chirag Gupta, Arun K Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Mahmoud Hegazy, Liviu Aolaritei, Michael Jordan, and Aymeric Dieuleveut. Valid selection among conformal sets. *Advances in Neural Information Processing Systems*, 38:174756–174787, 2026.

Thomas Holenstein. Parallel repetition: simplifications and the no-signaling case. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 411–419, 2007.

Max Hopkins, Russell Impagliazzo, Daniel Kane, Sihan Liu, and Christopher Ye. Replicability in high dimensional statistics. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1–8. IEEE, 2024.

Max Hopkins, Russell Impagliazzo, and Christopher Ye. Approximate replicability in learning. *arXiv preprint arXiv:2510.20200*, 2025.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. *arXiv preprint arXiv:2310.06430*, 2023.

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. One-shot federated conformal prediction. In *International Conference on Machine Learning*, pp. 14153–14177. PMLR, 2023.

Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In *Proceedings of the 54th annual ACM SIGACT symposium on theory of computing*, pp. 818–831, 2022.

MD Andras Janosi, MD Walter Steinbrunn, MD Matthias Pfisterer, DE Johann-Jakob Schmid, MD Sarbjit Sandhu, PhD Kern H Guppy, MS Stella Lee, and MD Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

Alkis Kalavasis, Amin Karbasi, Grigoris Velegkas, and Felix Zhou. On the computational landscape of replicable learning. *Advances in Neural Information Processing Systems*, 37:105887–105927, 2024.

Amin Karbasi, Grigoris Velegkas, Lin Yang, and Felix Zhou. Replicability in reinforcement learning. *Advances in Neural Information Processing Systems*, 36:74702–74735, 2023.

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.

Sihan Liu and Christopher Ye. Replicable uniformity testing. *Advances in Neural Information Processing Systems*, 37:32039–32075, 2024.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pp. 22942–22964. PMLR, 2023.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International conference on machine learning*, pp. 6765–6774. PMLR, 2020.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and iterate: Reducing prediction churn. *Advances in Neural Information Processing Systems*, 29, 2016.

Yinjie Min, Liuhua Peng, and Changliang Zou. Stable localized conformal prediction via transduction. *arXiv preprint arXiv:2605.01452*, 2026.

Eugene Ndiaye. Stable conformal prediction sets. In *International Conference on Machine Learning*, pp. 16462–16479. PMLR, 2022.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. Pac confidence sets for deep neural networks via calibrated prediction. *arXiv preprint arXiv:2001.00106*, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, pp. 27907–27947. PMLR, 2023.

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

Gordan Prastalo and Kevin Maik Jablonka. Reducing cross-sample prediction churn in scientific machine learning. *arXiv preprint arXiv:2605.13826*, 2026.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, volume 2024, pp. 11654–11681, 2024.

A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, pp. 2412–15115, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

Yan Scholten and Stephan Günnemann. Provably reliable conformal prediction sets in the presence of data poisoning. In *International Conference on Learning Representations*, volume 2025, pp. 55873–55897, 2025.

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

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

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

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

Appendix

Supplementary tables and figures are numbered S1, S2, ... and appear in Appendix F. Throughout the appendix, $p := 1 - \alpha$. A few letters (λ, G, Δ, r, k, J, U, H, I, Γ, η, b) carry a local meaning in some proofs, stated where they are introduced.

A Proofs for Section 3

Measurable formulation. Fix integers $d, n \geq 1$, a target miscoverage level $\alpha \in (0, 1)$, and a finite label set $\mathcal{Y}$ with $|\mathcal{Y}| = K \geq 2$. Equip $\mathcal{Y}$ and $2^{\mathcal{Y}}$ with their discrete sigma-fields. Let $\mathcal{P}_{\mathrm{ac}}$ be the class of all Borel probability laws on $\mathbb{R}^d \times \mathcal{Y}$ whose X-marginal is absolutely continuous with respect to d-dimensional Lebesgue measure. Let $(\mathsf{R}, \mathcal{R})$ and $(\mathsf{V}, \mathcal{V})$ be standard Borel spaces, let ν be a fixed probability law on R, and let $R \sim \nu$ be independent of the calibration sample and test pair. A procedure is a jointly measurable map

$$\mathcal{A} : (\mathbb{R}^d \times \mathcal{Y})^n \times \mathsf{R} \longrightarrow \mathsf{V}.$$

Each $v \in \mathsf{V}$ deterministically encodes a measurable deployed set map $\mathcal{C}_v : \mathbb{R}^d \to 2^{\mathcal{Y}}$, and

$$\Gamma(v, x, y) := \mathbf{1}\{y \in \mathcal{C}_v(x)\}$$

is jointly measurable. All randomness affecting the deployed map is included in R before v is output. No fresh randomness is drawn when $\mathcal{C}_v(x)$ is evaluated. Equality, list membership, and replication below concern the encoded output in V. If replication is instead intended to mean equality of deployed maps, assume explicitly that $v \mapsto \mathcal{C}_v$ is injective on the admissible output family.

For $\mathcal{D} \sim P^{\otimes n}$, $R \sim \nu$, and an independent test pair $(X, Y) \sim P$, define the marginal coverage

$$\mathrm{Cov}_P(\mathcal{A}) := \mathbb{P}\big\{Y \in \mathcal{C}_{\mathcal{A}(\mathcal{D};R)}(X)\big\},$$

where the probability averages over $\mathcal{D}$, R, and the independent test pair. The procedure is *uniformly two-sided ε-valid* (the main text's uniform ε-validity) if

$$\mathrm{Cov}_P(\mathcal{A}) \in [1 - \alpha - \varepsilon,\, 1 - \alpha + \varepsilon] \qquad \text{for every } P \in \mathcal{P}_{\mathrm{ac}}.$$

For $\delta \in [0, 1]$, the procedure is $(1, \delta)$*-list replicable* over $\mathcal{P}_{\mathrm{ac}}$ if, for every $P \in \mathcal{P}_{\mathrm{ac}}$, there is a deterministic set $L_P \subseteq \mathsf{V}$, depending on P but not on the realized sample or seed, such that

$$|L_P| \leq 1 \quad \text{and} \quad (P^{\otimes n} \otimes \nu)\{\mathcal{A}(\mathcal{D}; R) \in L_P\} \geq 1 - \delta.$$

For probability laws P, Q on a common measurable space $(\Omega, \mathcal{F})$, we use the convention $\mathrm{d_{TV}}(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$. When two runs share nothing (no seed and no public sample), their complete seeds are independent copies $R, R' \sim \nu$. Thus ρ-replicability without shared information means

$$\mathbb{P}\{\mathcal{A}(\mathcal{D}; R) \neq \mathcal{A}(\mathcal{D}'; R')\} \leq \rho \qquad \text{for every } P \in \mathcal{P}_{\mathrm{ac}},$$

where $\mathcal{D}, \mathcal{D}' \overset{\mathrm{iid}}{\sim} P^{\otimes n}$ and $\mathcal{D}, \mathcal{D}', R, R'$ are mutually independent. Shared-seed replication uses the same R in both runs.

A.1 Proof of Theorem 1

Proof. Suppose, toward a contradiction, that $\mathcal{A}$ is both $(1, \delta)$-list replicable and uniformly two-sided ε-valid, where $\varepsilon + \delta < \min\{\alpha, 1 - \alpha\}$. In particular, $\delta < 1/2$. Let

$$\mu_P := \mathcal{L}_{P^{\otimes n} \otimes \nu}(\mathcal{A}(\mathcal{D}; R)).$$

For each $P \in \mathcal{P}_{\mathrm{ac}}$, list replicability and $1 - \delta > 0$ imply that $L_P = \{v_P\}$ for some $v_P \in \mathsf{V}$ and

$$\mu_P(\{v_P\}) \geq 1 - \delta.$$

Because $1 - \delta > 1/2$, this heavy atom is unique.

We first show that v_P does not depend on P. By data processing, invariance of total variation under multiplication by the common seed law, and the product bound,

$$\begin{aligned}\mathrm{d_{TV}}(\mu_P, \mu_Q) &\le \mathrm{d_{TV}}(P^{\otimes n} \otimes \nu, Q^{\otimes n} \otimes \nu) \\ &= \mathrm{d_{TV}}(P^{\otimes n}, Q^{\otimes n}) \le 1 - \{1 - \mathrm{d_{TV}}(P, Q)\}^n \le n\ \mathrm{d_{TV}}(P, Q).\end{aligned}$$

Consequently, if $n\,\mathrm{d_{TV}}(P, Q) < 1 - 2\delta$, then

$$\mu_Q(\{v_P\}) \ge 1 - \delta - n\,\mathrm{d_{TV}}(P, Q) > \delta.$$

If $v_P \neq v_Q$, the corresponding singletons are disjoint, and

$$1 \ge \mu_Q(\{v_P\}) + \mu_Q(\{v_Q\}) > \delta + (1 - \delta) = 1,$$

a contradiction. Thus $v_P = v_Q$ whenever $n\,\mathrm{d_{TV}}(P, Q) < 1 - 2\delta$.

For arbitrary $P, Q \in \mathcal{P}_{\mathrm{ac}}$, put $P_t = (1 - t)P + tQ$, $0 \le t \le 1$. The class $\mathcal{P}_{\mathrm{ac}}$ is convex, and

$$\mathrm{d_{TV}}(P_s, P_t) = |s - t|\,\mathrm{d_{TV}}(P, Q) \le |s - t|.$$

Choose an integer m such that $n/m < 1 - 2\delta$. Applying the local argument successively to $P_{j/m}$, $j = 0, \ldots, m$, gives $v_P = v_Q$. Hence there is one $v^\star \in \mathsf{V}$ such that

$$\mu_P(\{v^\star\}) \ge 1 - \delta \qquad \text{for every } P \in \mathcal{P}_{\mathrm{ac}}.$$

For $v \in \mathsf{V}$, define

$$c_P(v) := \int_{\mathbb{R}^d \times \mathcal{Y}} \Gamma(v, x, y)\, P(dx, dy).$$

The function $v \mapsto c_P(v)$ is measurable by joint measurability of Γ and Tonelli's theorem. Independence of the test pair gives

$$\mathrm{Cov}_P(\mathcal{A}) = \int_{\mathsf{V}} c_P(v)\, \mu_P(dv).$$

Since $0 \le c_P \le 1$,

$$|\,\mathrm{Cov}_P(\mathcal{A}) - c_P(v^\star)| \le \int_{\mathsf{V} \setminus \{v^\star\}} |c_P(v) - c_P(v^\star)|\, \mu_P(dv) \le \delta.$$

Uniform two-sided validity therefore implies, for every $P \in \mathcal{P}_{\mathrm{ac}}$,

$$c_P(v^\star) \in [1 - \alpha - \varepsilon - \delta,\ 1 - \alpha + \varepsilon + \delta].$$

The assumed inequality makes the lower endpoint strictly positive and the upper endpoint strictly smaller than one.

For each $y \in \mathcal{Y}$, let

$$G_y := \{x \in \mathbb{R}^d : y \in \mathcal{C}_{v^\star}(x)\}.$$

Each G_y is Lebesgue measurable. If some G_y has positive Lebesgue measure, the Lebesgue density theorem supplies a density point $x_0 \in G_y$. For $h > 0$, let

$$P_{x_0,h,y} := \mathrm{Unif}(B(x_0, h)) \otimes \delta_y.$$

This law belongs to $\mathcal{P}_{\mathrm{ac}}$, and

$$c_{P_{x_0,h,y}}(v^\star) = \frac{\mathrm{Leb}(G_y \cap B(x_0, h))}{\mathrm{Leb}(B(x_0, h))} \longrightarrow 1 \qquad (h \downarrow 0),$$

contradicting the uniform strict upper bound. Hence every G_y is Lebesgue-null. Since $\mathcal{Y}$ is finite and P_X is absolutely continuous, for every $P \in \mathcal{P}_{\mathrm{ac}}$,

$$c_P(v^\star) = \sum_{y \in \mathcal{Y}} P\{X \in G_y, Y = y\} = 0,$$

contradicting the uniform strict lower bound. This proves the singleton-list impossibility.

For the consequence without shared information, fix $P \in \mathcal{P}_{\mathrm{ac}}$ and let $V = \mathcal{A}(\mathcal{D}; R)$ and $V' = \mathcal{A}(\mathcal{D}'; R')$. These variables are independent with common law μ_P. Because V is standard Borel,

its diagonal is measurable. If $\{v_j : j \in J_P\}$ is the at most countable set of point atoms of μ_P and $a_j := \mu_P(\{v_j\})$, Tonelli's theorem gives

$$\mathbb{P}\{V = V'\} = \int_{\mathsf{V}} \mu_P(\{v\})\, \mu_P(dv) = \sum_{j \in J_P} a_j^2.$$

If this collision probability is at least $1 - \rho > 0$, then J_P is nonempty. A summable family of positive masses has a largest member. If its positive supremum were not attained, infinitely many members would exceed half that supremum, contradicting summability. Write $a_P := \max_{j \in J_P} a_j$. Then

$$1 - \rho \le \sum_{j \in J_P} a_j^2 \le a_P \sum_{j \in J_P} a_j \le a_P.$$

Thus the singleton containing a largest atom captures the output with probability at least $1 - \rho$. Since this holds for every P, the procedure is $(1, \rho)$-list replicable. The final assertion follows by applying the first part with $\delta = \rho$. □

A.2 Atomless outputs

Lemma 1 (Atomless outputs never replicate)**.** *Let a calibrator output a scalar cutoff $V = \mathcal{A}(\mathcal{D}; r)$, and let μ_r be the conditional law of V given the shared randomness r. If μ_r is atomless for almost every r, then two independent runs $V = \mathcal{A}(\mathcal{D}; r)$ and $V' = \mathcal{A}(\mathcal{D}'; r)$ on the same r satisfy $\mathbb{P}(V = V') = 0$, and the deployed maps $\mathcal{C}_V$ and $\mathcal{C}_{V'}$ coincide only on the event that the score has no candidate label between V and V'. In particular, the bagged cutoff (the mean of the cutoffs of B bootstrap resamples), the median of the cutoffs of K folds, and a kernel-smoothed quantile all have threshold identity zero when F is continuous.*

Proof. Given r, V and V' are independent with the common law μ_r, so $\mathbb{P}(V = V' \mid r) = \int \mu_r(\{v\})\, \mu_r(dv) = 0$ when μ_r has no atoms, and integrating over r gives $\mathbb{P}(V = V') = 0$. Equal maps with $V \neq V'$ require that no candidate label of any input has its score in $(V \wedge V', V \vee V']$, which is the stated event. For the bagged cutoff, conditionally on the bootstrap indices the output is $\sum_i w_i(\pi) S_i$, where π is the rank pattern of the sample and, for each of the finitely many patterns, $w(\pi)$ is a fixed nonnegative weight vector that sums to one. Hence $\mathbb{P}(V = v) \le \sum_\pi \mathbb{P}(\sum_i w_i(\pi) S_i = v)$, and each term is zero because a fixed nontrivial linear combination of independent continuous variables has a continuous law. The fold median with an odd number of folds is one of the sample scores, so $\mathbb{P}(V = v) \le \sum_i \mathbb{P}(S_i = v) = 0$, and with an even number it is the mean of two of them, which has a continuous law by the argument for the bagged cutoff. The p-quantile of a kernel-smoothed empirical distribution function with a strictly positive kernel is continuous and strictly increasing in each score, so conditionally on the other scores it is a strictly increasing function of one score with the continuous law F, and its conditional law has no atoms. □

B Rounding in score units

This appendix states the results for Algorithm 1 with G the identity, where the width is a length in score units and the bounds depend on the local density. Here and in Appendices C to F, RECAL without qualification denotes this seeded score-unit form. Proofs are in Appendix C.

Assumption 1 (Local margin)**.** *There are $\Delta > 0$, $0 < f_{\min} \le f_{\max} < \infty$ with F admitting a density $f \in [f_{\min}, f_{\max}]$ on $I = [q - \Delta, q + \Delta]$. Write $\kappa = f_{\max}/f_{\min}$. The condition is local, and it concerns the law of the realized score: probability-scale scores do not satisfy it automatically, since the probability integral transform makes $F(S)$ uniform while S keeps its own law.*

Theorem 5 (Upper bound)**.** *Let F be continuous, let Assumption 1 hold, let $k = \lceil (1-\alpha)(n+1) \rceil \le n$, and let $n \ge 4/(f_{\min}\Delta)$. Two analysts run* RECAL *with shared (β, u), $u \sim \mathrm{Unif}[0, \beta)$, on independent samples of size n. Then: **(i)** $\mathbb{P}[\tau \neq \tau'] \le \dfrac{B_n}{f_{\min}\beta} + 4e^{-n f_{\min}^2 \Delta^2/2}$, where $B_n := \sqrt{2\alpha(1-\alpha)/n} + \sqrt{2}/n$ is the quantity B_n^+ of Section 2. Appendices A to E drop the superscript. **(ii)** for every fixed offset u, marginal coverage is $\ge k/(n+1) \ge 1 - \alpha$, with no margin condition. And for δ with $e_n(\delta) := \sqrt{\log(2/\delta)/(2n)} + 2/n \le f_{\min}(\Delta - \beta)$, with probability $\ge 1 - \delta$ over the sample and then* simultaneously for every $u \in [0, \beta)$, $1 - \alpha - e_n(\delta) \le F(\tau) \le 1 - \alpha + e_n(\delta) + f_{\max}\beta$. ***(iii)*** *on $\{\tau = \tau'\}$ the two deployed maps (same score) coincide at every test input. (Proof: Appendix C.)*

The replication bound (i) and the marginal claim in (ii) do not use $\beta \le \Delta/2$. Only the localized coverage band does, through the condition $e_n(\delta) \le f_{\min}(\Delta - \beta)$. The leading mismatch term is read off the exact $\mathrm{Beta}(k, n{+}1{-}k)$ law of $F(\widetilde{\tau}_n)$, and localization and the band use the DKW inequality (Massart, 1990).

Corollary 3 (The (ε, ρ) protocol)**.** *Given $(\varepsilon, \rho, \delta)$ with $\varepsilon \le f_{\max}\Delta$, set $\beta = \varepsilon/(2f_{\max})$ (so $\beta \le \Delta/2$). With the explicit finite-sample remainder n_0 of* (12) *in Appendix C (for fixed α, δ and margin constants, $n_0 = o((\varepsilon\rho)^{-2})$), $n \ge \max\{32\kappa^2\alpha(1-\alpha)/(\varepsilon^2\rho^2),\ 8\log(2/\delta)/\varepsilon^2,\ n_0\}$ makes* RECAL *ρ-replicable with marginal coverage $\ge 1-\alpha$ and, with probability $1-\delta$ over the calibration sample and for every realized shared offset, $F(\tau) \in [1-\alpha-\varepsilon/2,\ 1-\alpha+\varepsilon]$.*

Proposition 2 (The plug-in width: asymptotics without $\hat{\kappa}$)**.** *Fix $\alpha \in (0,1)$, $\rho \in (0,1]$, and a score law with density f continuous and positive at $q = F^{-1}(1-\alpha)$. Let $\hat{f}_n$ be any shared pilot with $\hat{f}_n \to_p f(q)$, $\hat{f}_n \ge f_0 > 0$, independent of both calibration samples, and run* RECAL *with $\beta_n = c_0 B_n/(\hat{f}_n\rho)$ for a constant $c_0 > 0$. The choice $c_0 = 1$ gives the plug-in width of Section 7 at $\hat{\kappa}{=}1$ up to the $\sqrt{2}/n$ term of B_n. Then, with Z standard normal, (i) $\mathbb{P}[\tau \ne \tau'] \to \mathbb{E}[\min(\rho|Z|/c_0, 1)] \le \rho\sqrt{2/\pi}/c_0$. (ii) $\mathbb{E}F(\tau) - \frac{k}{n+1} = (1+o(1))\, c_0 B_n/(2\rho)$. And (iii) $\mathbb{P}[F(\tau) \le \frac{k}{n+1} + e_n(\delta) + (c_0 + o(1))B_n/\rho] \ge 1 - \delta - o(1)$. The width, hence the premium, does not depend on κ: the factor κ^2 in Corollary 3 appears only in the finite-sample certificate and vanishes in the limit. The pilot density estimate is distribution-dependent shared information, so this limit is not a statement in the information regime of Theorem 4, and we do not compare its constants with Corollary 1. The choice $c_0 = \sqrt{2/\pi}$ makes the limit in (i) equal to ρ to first order as $\rho \to 0$, and strictly less at fixed ρ. At $c_0 = 1$ the limit is at most 0.80ρ. (Proof: Appendix C. Empirical check: Appendix F.)*

Proposition 3 (RECAL$^+$ attains the cap order)**.** *Let Assumption 1 hold with F continuous. Fix $\delta \in (0,1)$, put $p_+ := p + e_n(\delta)$ with $e_n(\delta) = \sqrt{\log(2/\delta)/(2n)} + 2/n$, and assume $e_n(\delta) \le f_{\min}\Delta/4$ and $k^+ := \lceil p_+(n+1)\rceil \le n$.* RECAL$^+$ *rounds $\widetilde{\tau}^+ := S_{(k^+)}$ upward on the grid $u + \beta\mathbb{Z}$ with a shared offset $u \sim \mathrm{Unif}[0,\beta)$ and $\beta \le \Delta/2$. Write $\alpha_+ := 1 - p_+$ and $B_n(\alpha_+) := \sqrt{2\alpha_+(1-\alpha_+)/n} + \sqrt{2}/n$. Then (i) for every offset, $\mathbb{P}[F(\tau^+) \ge 1-\alpha] \ge 1-\delta$. (ii) $\mathbb{P}[\tau_A^+ \ne \tau_B^+] \le B_n(\alpha_+)/(f_{\min}\beta) + 4\exp\{-2n(f_{\min}\Delta/2 - 2/n)_+^2\}$. (iii) With probability at least $1-\delta$ over the sample and simultaneously for every offset, $1-\alpha \le F(\tau^+) \le 1-\alpha + 2e_n(\delta) + f_{\max}\beta$. Hence, with $\beta := 2B_n(\alpha_+)/(f_{\min}\rho)$ and n large enough that $\beta \le \Delta/2$ and the tail in (ii) is at most $\rho/2$,* RECAL$^+$ *is ρ-replicable and satisfies the hypotheses of Corollary 1 with failure probability δ and cap $e = 2e_n(\delta) + 2\kappa B_n(\alpha_+)/\rho = O\big(\sqrt{\log(1/\delta)/n} + \kappa\sqrt{\alpha(1-\alpha)/n}/\rho\big)$, which matches the lower bound $\Omega(\sqrt{\alpha(1-\alpha)/n}/\rho)$ of the corollary up to the factor κ whenever $\rho\sqrt{\log(1/\delta)}$ is bounded. The constants are uniform over any class of laws in which Assumption 1 holds with common $(f_{\min}, f_{\max}, \Delta)$, in particular over the two-piece family $\mathcal{P}_2(\alpha - e/2, e/2)$ of Corollary 1, where $(f_{\min}, f_{\max}, \Delta) = (\min\{\alpha, 1-\alpha\}, 2, \frac{1}{3})$ can be used: for $e \le \min(\alpha, 1-\alpha)/4$ this family consists of the laws P_m with $m \in [p - e/2, p + 3e/2] \subseteq [p - \frac{1}{2}\min\{\alpha,p\}, p + \frac{1}{2}\min\{\alpha,p\}]$, the range covered by Step 1 of Appendix D at $(\alpha, \varepsilon) = (\alpha, \min\{\alpha,p\}/4)$, which gives $q_m \in [\frac{1}{3}, \frac{2}{3}]$ and a density in $[\min\{\alpha,p\}, 2]$ on $[0,1]$. So* RECAL$^+$ *with β set from these constants and $\delta \le \frac{1}{32}$ lies in the class of that corollary. (Proof: Appendix C.)*

Proposition 4 (Fixed seed, many analysts, shift)**.** *Condition on the public protocol (pilot and width β_n) and on the normalized phase $V := u/\beta_n$. Write $\pi(V) := \mathbb{P}[\tau \ne \tau' \mid V]$, $\pi_2 := \mathbb{E}_V\pi(V)$, $\sigma_{S,n} := \sigma_F/f(q)$ for the standard deviation of the raw cutoff in score units, and let $p_j(V)$ be the $\mathcal{N}(q, \sigma_{S,n}^2)$ mass of grid cell j. Under the conditions of Proposition 2, with width constant c_0: (i) $\mathbb{P}_V[\pi(V) > t] \le \pi_2/t$. (ii) $\pi(V) - \{1 - \sum_j p_j(V)^2\} \to 0$. Since $\beta_n/\sigma_{S,n} \to c_0\sqrt{2}/\rho$, within this Gaussian lattice the two cells adjacent to the boundary nearest q carry all but $2\Phi(-c_0/(\sqrt{2}\rho))$ of the mass (at $\rho{=}.1$, below 2×10^{-12} for $c_0 = 1$ and 10^{-44} for $c_0 = 2$), so $1 - \sum_j p_j(V)^2 = 2\Phi(z_V)\{1 - \Phi(z_V)\}$ up to that truncation, with $z_V = \ell(V)/\sigma_{S,n}$ and $\ell(V)$ the distance from q to the nearest boundary. The truncation error does not include the error of the normal approximation itself. For a fixed seed, the two analysts either almost never mismatch or mismatch often. (iii) Averaged over V, the probability that M analysts sharing (β_n, V) all coincide tends to $1 - \mathbb{E}\min\{\rho R_M/(c_0\sqrt{2}), 1\}$, where R_M is the range of M standard normals. Up to the truncation remainder $\mathbb{E}(\rho R_M/(c_0\sqrt{2}) - 1)_+$, which vanishes as $\rho\sqrt{\log M} \to 0$, this is $1 - (d_M/d_2)\pi_2$ with $d_M := \mathbb{E}R_M$. (iv) If one analyst's score law is a location shift by θ of the other's, $\mathbb{P}[\tau_A = \tau_B] \ge 1 - |\theta|/\beta_n - \pi_2 - o(1)$, with each*

analyst's own coverage unchanged. Without the location-shift assumption, the two fluctuation scales can differ. (Proofs: Appendix C.)

C PROOFS FOR SECTIONS 2, 4 AND 5

Conventions. Throughout the appendices, $F^{-1}(v) := \inf\{t \in \mathbb{R} : F(t) \geq v\}$, $p := 1 - \alpha$, and $k_n := \lceil (1-\alpha)(n+1) \rceil$, which is the manuscript's k. Scores are smaller-is-more-conforming: for a fixed score s, write $\mathcal{C}_t^s(x) := \{y \in \mathcal{Y} : s(x,y) \leq t\}$, suppressing the superscript when the score is clear, so that increasing the threshold can only enlarge the prediction set.

Measurable setup. Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be measurable spaces, and let P be a probability measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$. All score functions below are measurable with respect to the indicated sigma-fields. Write P_X for the $\mathcal{X}$-marginal of P.

Remark 1 (Continuity by deterministic jitter). Replace s by $\widetilde{s}(x,y) = s(x,y) - u(x)\, g(x,y)$ with fixed measurable $u : \mathcal{X} \to [0,1)$, $g > 0$, used identically by all analysts (adapted from Scholten & Günnemann, 2025): ties break identically and marginal validity is unaffected. Continuity of the induced law holds when $u(X)$ is atomless and independent of $(s(X,Y), g(X,Y))$, and more generally when its conditional law given that pair is atomless: given $(s,g) = (a,\gamma)$, the event $\widetilde{s} = t$ forces $u(X) = (a-t)/\gamma$, which is then a null event for almost every (a,γ). Marginal atomlessness alone is insufficient ($s = u = x$, $g \equiv 1$ gives $\widetilde{s} \equiv 0$), and a finite-range hash has atoms, so experiments with such a jitter do not instantiate the continuous theory exactly.

Proposition 5 (Standard split conformal never replicates, full statement)**.** *Let $n \geq 1$ and*

$$\alpha \in \left[\frac{1}{n+1}, 1\right), \qquad k := \lceil (n+1)(1-\alpha) \rceil \in \{1, \ldots, n\}.$$

Let $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a fixed measurable nonconformity score. Put $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$, and, for $t \in \overline{\mathbb{R}}$, define $\mathcal{C}_t(x) := \{y \in \mathcal{Y} : s(x,y) \leq t\}$. Thus $\mathcal{C}_{+\infty}(x) = \mathcal{Y}$ and $\mathcal{C}_{-\infty}(x) = \varnothing$ for every x. Suppose that $F(t) := \Pr\{s(X,Y) \leq t\}$, $(X,Y) \sim P$, is continuous. Let $\mathcal{D}_n := ((X_i, Y_i))_{i=1}^n$ and $\mathcal{D}'_n := ((X'_i, Y'_i))_{i=1}^n$ be independent i.i.d. calibration samples from P. Write $S_i := s(X_i, Y_i)$, $S'_i := s(X'_i, Y'_i)$, and let $\tau_n := S_{(k)}$, $\tau'_n := S'_{(k)}$ be the corresponding kth order-statistic thresholds. Then

$$F(\tau_n),\, F(\tau'_n) \overset{\text{ind}}{\sim} \mathrm{Beta}(k, n+1-k), \qquad \Pr\{\tau_n = \tau'_n\} = 0.$$

Let $(X,Y) \sim P$ be independent of both calibration samples and define the conditional operational disagreement probability

$$\Delta_n := \Pr\left\{ \mathbf{1}\{Y \in \mathcal{C}_{\tau_n}(X)\} \neq \mathbf{1}\{Y \in \mathcal{C}_{\tau'_n}(X)\} \,\middle|\, \mathcal{D}_n, \mathcal{D}'_n \right\}.$$

Then $\Delta_n = |F(\tau_n) - F(\tau'_n)| > 0$ almost surely. Define the following possibly nonmeasurable subsets of the joint calibration-sample space:

$$E_{\text{ae}} := \left\{ \begin{array}{l} \text{there exists } N \in \mathcal{A} \text{ with } P_X(N) = 0 \text{ such that} \\ \mathcal{C}_{\tau_n}(x) = \mathcal{C}_{\tau'_n}(x) \text{ for every } x \notin N \end{array} \right\},$$
$$E_{\text{pt}} := \{\mathcal{C}_{\tau_n}(x) = \mathcal{C}_{\tau'_n}(x) \text{ for every } x \in \mathcal{X}\}.$$

If $\Pr^*$ denotes outer probability under the joint law of the two calibration samples, then $\Pr^*(E_{\text{ae}}) = \Pr^*(E_{\text{pt}}) = 0$. Consequently, both subsets are measurable in the completion, where each has probability zero. Thus the prediction-set maps fail to replicate exactly almost surely. Finally, if $(X_{n+1}, Y_{n+1}) \sim P$ is independent of $\mathcal{D}_n$, then the calibration-conditional population coverage

$$\Gamma_n := \Pr\{Y_{n+1} \in \mathcal{C}_{\tau_n}(X_{n+1}) \,|\, \mathcal{D}_n\}$$

satisfies $\Gamma_n = F(\tau_n) \sim \mathrm{Beta}(k, n+1-k)$, and consequently $\Pr\{Y_{n+1} \in \mathcal{C}_{\tau_n}(X_{n+1})\} = k/(n+1) \geq 1 - \alpha$.

Proof. Continuity of F and the probability integral transform give

$$V_i := F(S_i) \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1), \qquad V'_i := F(S'_i) \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1).$$

The two collections are independent. Since F is nondecreasing, applying F commutes with taking order statistics, even if F has flat intervals:

$$F(\tau_n) = F(S_{(k)}) = V_{(k)}, \qquad F(\tau'_n) = F(S'_{(k)}) = V'_{(k)}$$

almost surely. The kth order statistic of n independent $\mathrm{Unif}(0,1)$ variables has density

$$f_k(v) = \frac{n!}{(k-1)!(n-k)!} v^{k-1}(1-v)^{n-k}, \qquad 0 < v < 1.$$

Therefore, $F(\tau_n), F(\tau'_n) \overset{\text{ind}}{\sim} \mathrm{Beta}(k, n+1-k)$. This Beta distribution is atomless, and hence $\Pr\{F(\tau_n) = F(\tau'_n)\} = 0$. Since $\{\tau_n = \tau'_n\} \subseteq \{F(\tau_n) = F(\tau'_n)\}$, it follows that $\Pr\{\tau_n = \tau'_n\} = 0$.

For deterministic $t, t' \in \mathbb{R}$, the two acceptance events are nested, so

$$\begin{aligned}\Pr\{\mathbf{1}\{Y \in \mathcal{C}_t(X)\} \neq \mathbf{1}\{Y \in \mathcal{C}_{t'}(X)\}\} &= \Pr\{\min(t,t') < s(X,Y) \leq \max(t,t')\} \\ &= |F(t) - F(t')|.\end{aligned}$$

Substituting the calibration-measurable thresholds and conditioning on the two independent calibration samples therefore yields $\Delta_n = |F(\tau_n) - F(\tau'_n)| > 0$ almost surely.

For any realized t, t', if $\mathcal{C}_t(x) = \mathcal{C}_{t'}(x)$ outside some $N \in \mathcal{A}$ with $P_X(N) = 0$, then the measurable disagreement set is contained in $N \times \mathcal{Y}$. Hence

$$\mathbf{1}\{Y \in \mathcal{C}_t(X)\} = \mathbf{1}\{Y \in \mathcal{C}_{t'}(X)\} \qquad P\text{-almost surely.}$$

The preceding disagreement identity then gives $F(t) = F(t')$. Consequently,

$$E_{\mathrm{pt}} \subseteq E_{\mathrm{ae}} \subseteq \{F(\tau_n) = F(\tau'_n)\}.$$

The event on the right is measurable and has probability zero. Both subsets on the left therefore have outer probability zero and belong to the completion with completed probability zero.

For the independent test observation (X_{n+1}, Y_{n+1}), $\Gamma_n = \Pr\{s(X_{n+1}, Y_{n+1}) \leq \tau_n \mid \mathcal{D}_n\} = F(\tau_n)$ almost surely. The asserted Beta law follows from the order-statistic calculation above. Taking expectations and using the mean of a $\mathrm{Beta}(k, n+1-k)$ random variable gives $\Pr\{Y_{n+1} \in \mathcal{C}_{\tau_n}(X_{n+1})\} = \mathbb{E}[\Gamma_n] = k/(n+1)$. Finally, $k = \lceil (n+1)(1-\alpha) \rceil \geq (n+1)(1-\alpha)$, so $k/(n+1) \geq 1-\alpha$. □

Order-statistic facts. For continuous F and $k_n = \lceil p(n+1) \rceil \leq n$,

$$U := F(\widetilde{\tau}) = F(S_{(k_n)}) \sim \mathrm{Beta}(k_n, n+1-k_n), \qquad m_n := \mathbb{E}U = \frac{k_n}{n+1}, \tag{1}$$

with

$$p \leq m_n < p + \frac{1}{n+1}, \qquad 0 \leq \frac{k_n}{n} - p < \frac{2}{n}, \qquad \mathrm{Var}(U) = \frac{m_n(1-m_n)}{n+2}. \tag{2}$$

Continuity implies that the sample scores have no ties almost surely, so $\widehat{F}_n(\widetilde{\tau}) = k_n/n$.

Lemma 2 (Mismatch under a uniform offset)**.** *For fixed $a, b \in \mathbb{R}$ and $u \sim \mathrm{Unif}[0, \beta)$,*

$$\mathbb{P}_u\{\mathrm{rd}_u(a) \neq \mathrm{rd}_u(b)\} = \frac{\min\{|a-b|, \beta\}}{\beta}. \tag{3}$$

Proof. The two round-ups differ exactly when a point of $u + \beta\mathbb{Z}$ lies in the half-open interval between a and b. If $d := |a-b| \geq \beta$, every grid phase has such a point. If $d < \beta$, reducing the phase modulo β shows that the set of offsets producing such a point is an arc of length d in $[0, \beta)$. Dividing its Lebesgue measure by β proves (3). □

Lemma 3 (Localization)**.** *Let F be continuous, let $S_1, \ldots, S_n$ be i.i.d. with cdf F, let $k_n = \lceil (1-\alpha)(n+1) \rceil \leq n$, and put $\widetilde{\tau} = S_{(k_n)}$. Under Assumption 1, $|F(\widetilde{\tau}) - p| < f_{\min}\Delta$ implies $\widetilde{\tau} \in I$, and for every $a, b \in I$,*

$$|a-b| \leq \frac{|F(a) - F(b)|}{f_{\min}}. \tag{4}$$

If in addition $n \geq 4/(f_{\min}\Delta)$, then

$$\mathbb{P}\{|F(\widetilde{\tau}) - p| \geq f_{\min}\Delta\} \leq 2\exp\{-2n(f_{\min}\Delta - 2/n)^2\} \leq 2\exp\left(-\frac{n f_{\min}^2 \Delta^2}{2}\right). \tag{5}$$

Proof. Continuity gives $F(q) = p$. Absolute continuity and the lower density bound give

$$F(q+\Delta) - p \geq f_{\min}\Delta, \qquad p - F(q-\Delta) \geq f_{\min}\Delta.$$

Monotonicity therefore proves the localization claim. If $a < b$ are in I, then

$$F(b) - F(a) = \int_a^b f(t)\,dt \geq f_{\min}(b-a),$$

which proves (4) without requiring pointwise differentiability of F. By (2), $|F(\widetilde{\tau}) - p| \leq \|\widehat{F}_n - F\|_\infty + 2/n$. The Dvoretzky–Kiefer–Wolfowitz inequality therefore gives the first bound in (5). Since $n \geq 4/(f_{\min}\Delta)$ implies $f_{\min}\Delta - 2/n \geq f_{\min}\Delta/2$, the second bound follows. □

PROOF OF THEOREM 5

Let E be the event that both unrounded thresholds satisfy $|F(\widetilde{\tau}_j) - p| < f_{\min}\Delta$. Lemma 3 and a union bound give

$$\mathbb{P}(E^c) \leq 4\exp\left(-\frac{n f_{\min}^2 \Delta^2}{2}\right). \tag{6}$$

Conditional on the two unrounded thresholds, Lemma 2, followed on E by (4), gives

$$\begin{aligned}\mathbb{P}(\tau_A \neq \tau_B) &\leq \mathbb{E}\left[\frac{\min\{|\widetilde{\tau}_A - \widetilde{\tau}_B|, \beta\}}{\beta}\mathbf{1}_E\right] + \mathbb{P}(E^c) \\ &\leq \frac{\mathbb{E}|F(\widetilde{\tau}_A) - F(\widetilde{\tau}_B)|}{f_{\min}\beta} + 4\exp\left(-\frac{n f_{\min}^2\Delta^2}{2}\right).\end{aligned} \tag{7}$$

The transformed thresholds are independent copies U, U' of the beta variable in (1), and hence

$$\mathbb{E}|U - U'| \leq \sqrt{\mathbb{E}(U-U')^2} = \sqrt{2\,\mathrm{Var}(U)}. \tag{8}$$

Writing $m_n = p + r_n$ with $0 \leq r_n < 1/(n+1)$,

$$m_n(1-m_n) = \alpha(1-\alpha) + r_n(2\alpha - 1) - r_n^2 \leq \alpha(1-\alpha) + \frac{\alpha}{n+1}.$$

Consequently,

$$\mathrm{Var}(U) \leq \frac{\alpha(1-\alpha)}{n} + \frac{1}{n^2}, \tag{9}$$

and

$$\mathbb{E}|U - U'| \leq \sqrt{\frac{2\alpha(1-\alpha)}{n} + \frac{2}{n^2}} \leq \sqrt{\frac{2\alpha(1-\alpha)}{n}} + \frac{\sqrt{2}}{n}.$$

Substitution into (7) proves part (i).

For marginal validity, upward rounding gives $\widetilde{\tau} \leq \tau < \widetilde{\tau} + \beta$ for every offset, so $\mathcal{C}_{\widetilde{\tau}}(x) \subseteq \mathcal{C}_\tau(x)$ for every x. Among $n+1$ i.i.d. continuous scores, the test score has a uniform rank. Therefore

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}_\tau(X_{n+1})\} \geq \mathbb{P}\{S_{n+1} \leq S_{(k_n)}\} = \frac{k_n}{n+1} \geq 1 - \alpha,$$

proving the marginal claim of part (ii). This argument holds for every fixed offset and does not use the margin condition.

For the training-conditional statement, DKW implies that, with probability at least $1-\delta$,

$$|F(\widetilde{\tau}) - p| \leq \|\widehat{F}_n - F\|_\infty + \left|\frac{k_n}{n} - p\right| \leq e_n(\delta). \tag{10}$$

Under the condition $e_n(\delta) \leq f_{\min}(\Delta - \beta)$, this first localizes $\widetilde{\tau}$ to I and then, by (4), gives

$$|\widetilde{\tau} - q| \leq \frac{e_n(\delta)}{f_{\min}} \leq \Delta - \beta.$$

Thus $[\widetilde{\tau}, \widetilde{\tau} + \beta] \subseteq I$. Monotonicity and upward rounding give $F(\tau) \geq F(\widetilde{\tau}) \geq p - e_n(\delta)$, while absolute continuity and the upper density bound give

$$F(\tau) \leq F(\widetilde{\tau}) + f_{\max}(\tau - \widetilde{\tau}) \leq p + e_n(\delta) + f_{\max}\beta.$$

The DKW event depends only on the sample, so these inequalities hold simultaneously for every offset. This proves part (ii).

Finally, if $\tau_A = \tau_B = \tau$ and the score map is shared, then both analysts deploy $\mathcal{C}_\tau^s(x) = \{y : s(x,y) \leq \tau\}$ for every x, which proves part (iii). □

Full statement of Corollary 3. Let s be a fixed measurable score map whose score cdf F under P is continuous and satisfies Assumption 1. The two analysts use independent i.i.d. calibration samples from $P^{\otimes n}$. Fix $\alpha, \varepsilon, \rho, \delta \in (0,1)$ and assume

$$\varepsilon \le f_{\max}\Delta. \tag{11}$$

Set

$$\beta := \frac{\varepsilon}{2f_{\max}}, \qquad h := f_{\min}(\Delta - \beta) > 0, \qquad n_\alpha := \left\lfloor \frac{1}{\alpha} \right\rfloor,$$

and define the explicit finite-sample remainder

$$n_0 := \left\lceil \max\left\{ n_\alpha, \frac{8}{\varepsilon}, \frac{8\sqrt{2}\,\kappa}{\varepsilon\rho}, \frac{2\log(16/\rho)}{f_{\min}^2\Delta^2}, \frac{2\log(2/\delta)}{h^2}, \frac{4}{h} \right\} \right\rceil. \tag{12}$$

If

$$n \ge \max\left\{ \frac{32\kappa^2\alpha(1-\alpha)}{\varepsilon^2\rho^2}, \frac{8\log(2/\delta)}{\varepsilon^2}, n_0 \right\}, \tag{13}$$

then two analysts running RECAL on independent samples of size n, with the same score and the independent shared offset $u \sim \mathrm{Unif}[0,\beta)$, satisfy $\mathbb{P}_{\mathcal{D},\mathcal{D}',u}(\tau \ne \tau') \le \rho$. Consequently, their deployed prediction-set maps are ρ-replicable. Each analyst has marginal coverage at least $1-\alpha$ and, with probability at least $1-\delta$ over its calibration sample,

$$F(\tau) \in \left[1 - \alpha - \frac{\varepsilon}{2},\ 1 - \alpha + \varepsilon\right]. \tag{14}$$

On the event giving (14), the inclusion holds for every realized shared offset u.

PROOF OF COROLLARY 3

Condition (11) gives $\beta \le \Delta/2$. Also, $n \ge n_\alpha$ implies $\alpha > 1/(n+1)$ and hence $k_n \le n$, while $n \ge 4/h \ge 4/(f_{\min}\Delta)$ gives the localization sample-size condition of Theorem 5. Using $\beta = \varepsilon/(2f_{\max})$ in Theorem 5(i) gives

$$\mathbb{P}(\tau \ne \tau') \le \frac{2\kappa}{\varepsilon}\sqrt{\frac{2\alpha(1-\alpha)}{n}} + \frac{2\sqrt{2}\,\kappa}{\varepsilon n} + 4\exp\left(-\frac{n f_{\min}^2\Delta^2}{2}\right). \tag{15}$$

The first lower bound on n in (13) makes the first term at most $\rho/2$. The conditions

$$n \ge \frac{8\sqrt{2}\,\kappa}{\varepsilon\rho}, \qquad n \ge \frac{2\log(16/\rho)}{f_{\min}^2\Delta^2}$$

make the second and third terms at most $\rho/4$ each. Thus $\mathbb{P}(\tau \ne \tau') \le \rho$, and equality of the deployed maps on the complement follows from the shared score. For the conditional coverage band,

$$n \ge \frac{8\log(2/\delta)}{\varepsilon^2}, \qquad n \ge \frac{8}{\varepsilon}$$

imply, respectively, $\sqrt{\log(2/\delta)/(2n)} \le \varepsilon/4$ and $2/n \le \varepsilon/4$, so $e_n(\delta) \le \varepsilon/2$. Likewise,

$$n \ge \frac{2\log(2/\delta)}{h^2}, \qquad n \ge \frac{4}{h}$$

give $e_n(\delta) \le h = f_{\min}(\Delta - \beta)$. Theorem 5(ii) and $f_{\max}\beta = \varepsilon/2$ now yield

$$1 - \alpha - \frac{\varepsilon}{2} \le F(\tau) \le 1 - \alpha + \frac{\varepsilon}{2} + f_{\max}\beta = 1 - \alpha + \varepsilon.$$

Marginal coverage follows from Theorem 5(ii). For the order of n_0: since $\beta \le \Delta/2$ gives $h \ge f_{\min}\Delta/2$, every entry of (12) is $O(1/(\varepsilon\rho) + \log(1/\rho))$ for fixed α, δ and margin constants, so $n_0 = o((\varepsilon\rho)^{-2})$ as $\varepsilon\rho \to 0$. □

Corollary 4 (Unequal calibration sizes)**.** *Analysts sharing (s, α, β, u) with sizes $n_A \ne n_B$, where $\alpha \ge \max\{1/(n_A+1), 1/(n_B+1)\}$ and $n_A, n_B \ge 4/(f_{\min}\Delta)$, satisfy, for* every $\beta > 0$, $\mathbb{P}[\tau_A \ne \tau_B] \le \left[\sqrt{\alpha(1-\alpha)}(n_A^{-1/2} + n_B^{-1/2}) + 3/(n_A \wedge n_B)\right]/(f_{\min}\beta) + 4e^{-(n_A \wedge n_B) f_{\min}^2 \Delta^2/2}$. *Each analyst separately retains marginal coverage $\ge 1-\alpha$, and per-analyst conditional bands hold as in Theorem 5(ii) with a union bound (full statement below).*

Full statement of Corollary 4. Suppose both analysts use the same score s, whose continuous score distribution F satisfies Assumption 1. Let their calibration samples be independent i.i.d. samples from the same population P, with sizes n_A, n_B, and assume

$$\alpha \geq \max\left\{\frac{1}{n_A+1}, \frac{1}{n_B+1}\right\}, \qquad n_A, n_B \geq \frac{4}{f_{\min}\Delta}.$$

For any $\beta > 0$, let the analysts use the same independent shared offset $u \sim \mathrm{Unif}[0, \beta)$ and their respective conformal ranks $k_j := \lceil (1-\alpha)(n_j+1) \rceil$, $j \in \{A, B\}$. Writing $m := n_A \wedge n_B$, their rounded thresholds satisfy

$$\mathbb{P}(\tau_A \neq \tau_B) \leq \frac{\sqrt{\alpha(1-\alpha)}(n_A^{-1/2} + n_B^{-1/2}) + 3/m}{f_{\min}\beta} + 4\exp\left(-\frac{m f_{\min}^2 \Delta^2}{2}\right). \tag{16}$$

Thus the analysts need not coordinate sample sizes. On $\{\tau_A = \tau_B\}$ their deployed set maps coincide everywhere, and each analyst separately retains marginal coverage at least $1-\alpha$. No restriction $\beta \leq \Delta/2$ is needed for (16). If, in addition, $\delta_A, \delta_B \in (0,1)$, $\delta_A + \delta_B < 1$, and $e_{n_j}(\delta_j) \leq f_{\min}(\Delta - \beta)$ for $j \in \{A, B\}$, then with probability at least $1 - \delta_A - \delta_B$ the following inequalities hold simultaneously for $j \in \{A, B\}$:

$$1 - \alpha - e_{n_j}(\delta_j) \leq F(\tau_j) \leq 1 - \alpha + e_{n_j}(\delta_j) + f_{\max}\beta.$$

PROOF OF COROLLARY 4

For $j \in \{A, B\}$, let

$$U_j := F(\widetilde{\tau}_j) \sim \mathrm{Beta}(k_j, n_j + 1 - k_j), \qquad \mu_j := \mathbb{E}U_j = \frac{k_j}{n_j+1}.$$

The variables U_A, U_B are independent. The ceiling definition gives

$$0 \leq \mu_j - p < \frac{1}{n_j+1}, \qquad |\mu_A - \mu_B| \leq \frac{1}{m+1}.$$

The variance calculation in (9), with n_j in place of n, gives $\mathrm{sd}(U_j) \leq \sqrt{\alpha(1-\alpha)/n_j} + 1/n_j$. Therefore

$$\begin{aligned}
\mathbb{E}|U_A - U_B| &\leq \mathbb{E}|U_A - \mu_A| + |\mu_A - \mu_B| + \mathbb{E}|U_B - \mu_B| \\
&\leq \sqrt{\alpha(1-\alpha)}(n_A^{-1/2} + n_B^{-1/2}) + \frac{1}{n_A} + \frac{1}{n_B} + \frac{1}{m+1} \\
&\leq \sqrt{\alpha(1-\alpha)}(n_A^{-1/2} + n_B^{-1/2}) + \frac{3}{m}.
\end{aligned}$$

Let E be the event that both raw thresholds localize to $I = [q - \Delta, q + \Delta]$. Applying Lemma 3 separately at the two sample sizes gives

$$\mathbb{P}(E^c) \leq 2e^{-n_A f_{\min}^2 \Delta^2/2} + 2e^{-n_B f_{\min}^2 \Delta^2/2} \leq 4e^{-m f_{\min}^2 \Delta^2/2}.$$

On E, $|\widetilde{\tau}_A - \widetilde{\tau}_B| \leq |U_A - U_B|/f_{\min}$. Conditioning on the raw thresholds and applying Lemma 2 now gives

$$\mathbb{P}(\tau_A \neq \tau_B) \leq \frac{\mathbb{E}|U_A - U_B|}{f_{\min}\beta} + 4e^{-m f_{\min}^2 \Delta^2/2},$$

which is (16). The shared score gives identical prediction-set maps whenever the thresholds agree. Apply Theorem 5(ii) separately to the two analysts for the marginal and conditional coverage claims, and use a union bound for the simultaneous claim. □

Corollary 5 (Robustness to imperfect sharing)**.** *Let analysts A, B use fixed measurable scores s_A, s_B on the same domain with $\|s_A - s_B\|_\infty \leq \eta_s$, and independent size-$n$ calibration samples from P_A, P_B. Let F, G be the CDFs of the* common *score $s_A(X,Y)$ under P_A and P_B (then $\mathrm{d_{TV}}(P_A, P_B) \leq \eta_P$ implies $\|F - G\|_\infty \leq \eta_P$, which fails if G is instead defined through s_B), both continuous, with Assumption 1 for F, effective margin $\lambda := f_{\min}\Delta - \eta_P > 0$, $k \leq n$, and $n \geq 4/\lambda$. With shared (β, u), $\mathbb{P}(\tau_A \neq \tau_B) \leq \left[\, b_n/(f_{\min}\beta) + (\eta_s + \eta_P/f_{\min})/\beta + 4e^{-n\lambda^2/2} \,\right] \wedge 1$,*

where $b_n := \sqrt{2\alpha(1-\alpha)/n + 2/n^2} \le B_n$: population mismatch shrinks the localization margin from $f_{\min}\Delta$ to λ. Each analyst separately retains its own marginal coverage $\ge 1-\alpha$. Equal numerical *thresholds under different scores need not give equal maps: under the window condition $e_n(\delta)/f_{\min} + \beta + \eta_s \le \Delta$, on a $1-\delta$ event, whenever $\tau_A = \tau_B = \tau$ the realized test pair is classified differently with probability $\le \min\{1, 2f_{\max}\eta_s\}$ (full-map version in the appendix). The condition $\eta_s = 0$ is sufficient, but not necessary, for map equality. This motivates distributing the score function as a hash-verified artifact. (Full statement and proof: Appendix C.)*

Full statement of Corollary 5. Let analysts A and B use independent calibration samples of size n from P_A and P_B, respectively. Let their score functions $s_A, s_B : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be fixed measurable maps on the same domain. Let $\eta_s, \eta_P \ge 0$, and assume

$$\|s_A - s_B\|_\infty \le \eta_s. \tag{17}$$

Define the common-score cdfs

$$F(t) := P_A\{s_A(X,Y) \le t\}, \qquad G(t) := P_B\{s_A(X,Y) \le t\}. \tag{18}$$

Assume F and G are continuous and

$$\|F - G\|_\infty \le \eta_P. \tag{19}$$

Condition (19) follows, for example, from $\mathrm{d_{TV}}(P_A, P_B) \le \eta_P$, because the same fixed score s_A appears in both cdfs. It would not, in general, follow from total variation if G were instead defined using s_B. Let $p := 1-\alpha$, $q := F^{-1}(p)$, and suppose F is absolutely continuous on $I = [q-\Delta, q+\Delta]$, with density satisfying $f_{\min} \le f \le f_{\max}$ almost everywhere on I. Define

$$\lambda := f_{\min}\Delta - \eta_P > 0, \qquad b_n := \sqrt{\frac{2\alpha(1-\alpha)}{n} + \frac{2}{n^2}}, \tag{20}$$

and assume $k_n = \lceil p(n+1) \rceil \le n$ and $n \ge 4/\lambda$. If both analysts use the same grid width $\beta > 0$ and the same independent offset $u \sim \mathrm{Unif}[0, \beta)$, then their numerical rounded thresholds satisfy

$$\mathbb{P}(\tau_A \ne \tau_B) \le \left[\frac{b_n}{f_{\min}\beta} + \frac{\eta_s + \eta_P/f_{\min}}{\beta} + 4\exp\Big\{ -\frac{n}{2}(f_{\min}\Delta - \eta_P)^2 \Big\} \right] \wedge 1. \tag{21}$$

Thus population mismatch reduces the effective localization margin from $f_{\min}\Delta$ to $f_{\min}\Delta - \eta_P$. Each analyst separately retains split-conformal marginal coverage at least $1-\alpha$ for a fresh test observation from its own population. Bound (21) concerns equality of the numerical thresholds. When $s_A \ne s_B$, equal thresholds need not define equal set maps.

Now let a fresh test pair $(X, Y) \sim P_A$ be independent of the entire calibration experiment. If, for some $\delta \in (0,1)$,

$$\frac{e_n(\delta)}{f_{\min}} + \beta + \eta_s \le \Delta, \tag{22}$$

then, with probability at least $1-\delta$ over the calibration experiment, every realization for which $\tau_A = \tau_B =: \tau$ satisfies

$$\mathbb{P}_{(X,Y)\sim P_A}(\mathbf{1}\{s_A(X,Y) \le \tau\} \ne \mathbf{1}\{s_B(X,Y) \le \tau\} \,|\, \mathcal{D}_A, \mathcal{D}_B, u) \le \min\{1, 2f_{\max}\eta_s\}. \tag{23}$$

Consequently,

$$\mathbb{P}(\tau_A = \tau_B,\ \mathbf{1}\{s_A(X,Y) \le \tau_A\} \ne \mathbf{1}\{s_B(X,Y) \le \tau_B\}) \le \min\{1, \delta + 2f_{\max}\eta_s\}. \tag{24}$$

If $\mathcal{Y}$ is finite and, for every $y \in \mathcal{Y}$, the cdf of $s_A(X, y)$ under $X \sim P_{A,X}$ is absolutely continuous on I with density bounded above by L_y almost everywhere, then on the same high-probability event and whenever $\tau_A = \tau_B = \tau$,

$$\mathbb{P}_{X\sim P_{A,X}}\{\mathcal{C}_\tau^{s_A}(X) \ne \mathcal{C}_\tau^{s_B}(X)\} \le \min\left\{1, 2\eta_s \sum_{y\in\mathcal{Y}} L_y\right\}. \tag{25}$$

The condition $\eta_s = 0$ is sufficient, but not necessary, for exact equality of the two set maps at a common threshold. The exact condition is $\mathbf{1}\{s_A(x,y) \le \tau\} = \mathbf{1}\{s_B(x,y) \le \tau\}$ for every (x, y).

PROOF OF COROLLARY 5

Let $\widetilde{\tau}_A$ be the k_nth order statistic of the s_A-scores in analyst A's sample. On analyst B's sample, let $\widetilde{\tau}_B$ be the k_nth order statistic of the s_B-scores and let $\widetilde{\tau}_B^A$ be the k_nth order statistic obtained by applying s_A to that same sample. An order statistic is 1-Lipschitz in the ℓ_∞ norm of the sample vector, so (17) gives

$$|\widetilde{\tau}_B - \widetilde{\tau}_B^A| \le \eta_s. \tag{26}$$

By continuity and the probability integral transform, $U := F(\widetilde{\tau}_A)$ and $V := G(\widetilde{\tau}_B^A)$ are independent copies of $\mathrm{Beta}(k_n, n+1-k_n)$. Define $E := \{|U-p| < \lambda,\ |V-p| < \lambda\}$. For either transformed order statistic, DKW and $|k_n/n - p| < 2/n$ give

$$\mathbb{P}(|U-p| \ge \lambda) \le 2\exp\{-2n(\lambda - 2/n)^2\} \le 2\exp(-n\lambda^2/2),$$

where the last inequality uses $n \ge 4/\lambda$. The same bound holds for V. Thus

$$\mathbb{P}(E^c) \le 4\exp(-n\lambda^2/2). \tag{27}$$

On E, $|F(\widetilde{\tau}_A) - p| < \lambda \le f_{\min}\Delta$, so $\widetilde{\tau}_A \in I$. Also,

$$|F(\widetilde{\tau}_B^A) - p| \le |G(\widetilde{\tau}_B^A) - p| + \|F - G\|_\infty < \lambda + \eta_P = f_{\min}\Delta,$$

so $\widetilde{\tau}_B^A \in I$. The inverse-Lipschitz inequality for F gives

$$|\widetilde{\tau}_A - \widetilde{\tau}_B^A| \le \frac{|F(\widetilde{\tau}_A) - F(\widetilde{\tau}_B^A)|}{f_{\min}} \le \frac{|U-V| + \eta_P}{f_{\min}}. \tag{28}$$

Combining (26) and (28) yields, on E,

$$|\widetilde{\tau}_A - \widetilde{\tau}_B| \le \frac{|U-V|}{f_{\min}} + \frac{\eta_P}{f_{\min}} + \eta_s.$$

Conditioning on the raw thresholds and applying the offset-collision identity therefore gives

$$\mathbb{P}(\tau_A \ne \tau_B) \le \frac{\mathbb{E}|U-V|}{f_{\min}\beta} + \frac{\eta_s + \eta_P/f_{\min}}{\beta} + 4e^{-n\lambda^2/2}.$$

Finally, $\mathbb{E}|U-V| \le \sqrt{2\,\mathrm{Var}(U)} \le b_n$ by (9), proving (21). For completeness, continuity of analyst B's own s_B-score cdf is not needed for marginal validity. Append independent continuous tie breakers to the $n+1$ exchangeable scores and rank the resulting pairs lexicographically. The test pair has a uniform rank. If its score is strictly larger than the k_nth calibration score, then its tie-broken rank is larger than k_n. Hence $\mathbb{P}\{S_{n+1} \le S_{(k_n)}\} \ge k_n/(n+1) \ge 1-\alpha$. Upward rounding can only enlarge the prediction set. Applying this argument to each analyst's own score and population proves both marginal-coverage claims.

On the analyst-A DKW event of probability at least $1-\delta$,

$$|F(\widetilde{\tau}_A) - p| \le e_n(\delta), \qquad |\widetilde{\tau}_A - q| \le \frac{e_n(\delta)}{f_{\min}}.$$

Since $\widetilde{\tau}_A \le \tau_A < \widetilde{\tau}_A + \beta$, condition (22) implies

$$[\tau_A - \eta_s, \tau_A + \eta_s] \subseteq I. \tag{29}$$

If $\tau_A = \tau_B = \tau$ and the two membership indicators differ, then $s_A(X,Y)$ and $s_B(X,Y)$ straddle τ. By (17), this forces $|s_A(X,Y) - \tau| \le \eta_s$. The upper density bound on the random interval in (29) therefore gives

$$P_A\{|s_A(X,Y) - \tau| \le \eta_s\} \le 2f_{\max}\eta_s,$$

which proves (23). Integrating over the exceptional DKW event proves (24). Finally, a union bound over labels, followed by the corresponding per-label density bound, proves (25). □

Proposition 6 (Pilot rule with a confidence band)**.** *Split a public pilot of $n_p = 2m$ scores (disjoint from all calibration data): $\hat{q}_p$ is the conformal quantile of the first half, N_W counts second-half scores in $W = [\hat{q}_p - h, \hat{q}_p + h]$ with $h \le \Delta/2$, and $\hat{f}_{\mathrm{L}} := \max\{N_W, 1\}/\big((1+c)\,2hm\big)$ for an accuracy $c \in (0,1)$. If $e_m(\delta_p/3) \le f_{\min}\Delta/2$ and $m \ge 3\log(3/\delta_p)/(2c^2 h f_{\min})$, then on a pilot event $\mathcal{G}_p$ of probability $\ge 1-\delta_p$: $W \subseteq I$ and $\frac{1-c}{1+c} f_{\min} \le \hat{f}_{\mathrm{L}} \le f_{\max}$. Set $\beta := 2\hat{\kappa}B_n/(\hat{f}_{\mathrm{L}}\rho)$, B_n as*

in Theorem 5, with shared offset $u = \beta V$, $V \sim \mathrm{Unif}[0,1)$. For every pilot realization in $\mathcal{G}_p$, under the explicit side conditions of the full statement below ($\overline{\beta}_n \le \Delta/2$, tail $\le \rho/2$), two analysts satisfy $\mathbb{P}[\tau \neq \tau' \mid \mathcal{D}_p] \le \rho$. Unconditionally $\mathbb{P}[\tau \neq \tau'] \le \rho + (1-\rho)\delta_p \le \rho + \delta_p$, and the middle expression equals a target $\rho_\star$ when the construction is run at $\rho_0 = (\rho_\star - \delta_p)/(1-\delta_p)$ (for $0 < \delta_p < \rho_\star$). Marginal coverage $\ge 1-\alpha$ holds for every *pilot and offset realization, and the conditional band holds with rounding excess at most $\frac{2(1+c)}{1-c}\hat{\kappa}\kappa B_n/\rho$. A valid $\hat{\kappa} \ge \kappa$ remains the one input that cannot be certified from data, and $\hat{\kappa} = 1.5$ kept mismatch at or below the target in the independent safety streams of Appendix F (digits and the Pythia sweep), except for Pythia-1.4B ($.103 \pm .018$ at ρ=.1). (Full statement and proof: Appendix C.)*

Full statement of Proposition 6. Suppose the standing continuous-score setup and Assumption 1 hold. Let the public pilot $\mathcal{D}_p$ consist of $n_p = 2m$ i.i.d. scores with distribution F, independent of all calibration samples, test observations, and shared random seeds. Fix

$$0 < h \le \frac{\Delta}{2}, \qquad c, \delta_p, \rho \in (0,1), \qquad k_p := \lceil (1-\alpha)(m+1) \rceil \le m.$$

Split $\mathcal{D}_p$ into two samples of size m. Let $\widehat{q}_p$ be the k_p-th order statistic of the first half and define

$$W := [\widehat{q}_p - h, \widehat{q}_p + h], \qquad N_W := \#\{S_i \text{ in the second pilot half} : S_i \in W\},$$

and the safeguarded estimator

$$\widehat{f}_{\mathrm{L}} := \frac{\max\{N_W, 1\}}{(1+c)\,2hm}.$$

For $j \ge 1$ and $\eta \in (0,1)$ write $e_j(\eta) := \sqrt{\log(2/\eta)/(2j)} + 2/j$. Assume

$$e_m(\delta_p/3) \le \frac{f_{\min}\Delta}{2}, \qquad m \ge \frac{3\log(3/\delta_p)}{2c^2 h f_{\min}}. \tag{30}$$

Then there is an event $\mathcal{G}_p \in \sigma(\mathcal{D}_p)$ with $\mathbb{P}(\mathcal{G}_p) \ge 1 - \delta_p$ such that, on $\mathcal{G}_p$,

$$W \subseteq I, \qquad N_W \ge 1, \qquad \frac{1-c}{1+c} f_{\min} \le \widehat{f}_{\mathrm{L}} \le f_{\max}. \tag{31}$$

Let $\widehat{\kappa} \ge \kappa = f_{\max}/f_{\min}$ be a pre-specified deterministic protocol input and set

$$B_n := \sqrt{\frac{2\alpha(1-\alpha)}{n}} + \frac{\sqrt{2}}{n}, \qquad \beta := \frac{2\widehat{\kappa} B_n}{\widehat{f}_{\mathrm{L}}\rho}.$$

Let $V \sim \mathrm{Unif}[0,1)$ be shared by the analysts and independent of the pilot, calibration samples, and test observations, and set $u = \beta V$. On $\mathcal{G}_p$,

$$\frac{2B_n}{f_{\min}\rho} \le \beta \le \overline{\beta}_n := \frac{2(1+c)}{1-c}\,\frac{\widehat{\kappa} B_n}{f_{\min}\rho}. \tag{32}$$

Let $k_n = \lceil (1-\alpha)(n+1) \rceil$ and suppose additionally that

$$k_n \le n, \qquad n \ge \frac{4}{f_{\min}\Delta}, \qquad \overline{\beta}_n \le \frac{\Delta}{2}, \qquad 4\exp\left(-\frac{n f_{\min}^2 \Delta^2}{2}\right) \le \frac{\rho}{2}. \tag{33}$$

For every pilot realization in $\mathcal{G}_p$, two analysts using that same public pilot and the same shared V, but independent calibration samples $\mathcal{D}, \mathcal{D}'$ of size n, satisfy

$$\mathbb{P}_{\mathcal{D},\mathcal{D}',V}[\tau(\mathcal{D}) \neq \tau(\mathcal{D}') \,|\, \mathcal{D}_p] \le \rho. \tag{34}$$

The probability in (34) averages over the shared random offset V. No bound conditional on a fixed realization of V is asserted. By Theorem 5(iii), the same upper bound applies to nonidentity of the two deployed prediction-set maps. If the random pilot is also included in the probability space, then

$$\mathbb{P}(\tau \neq \tau') \le \rho + (1-\rho)\delta_p \le \rho + \delta_p. \tag{35}$$

In particular, to guarantee an unconditional target $\rho_\star \in (0,1)$, choose $0 < \delta_p < \rho_\star$ and replace ρ throughout the construction and conditions above by $\rho_0 := (\rho_\star - \delta_p)/(1-\delta_p)$. Then the right-hand

side of (35) equals $\rho_\star$. For every realization of $\mathcal{D}_p$ and V, including pilot realizations outside $\mathcal{G}_p$, upward rounding retains the conditional-on-protocol marginal conformal guarantee

$$\mathbb{P}_{\mathcal{D},(X_{n+1},Y_{n+1})}[Y_{n+1} \in \mathcal{C}_\tau(X_{n+1}) \,|\, \mathcal{D}_p, V] \geq 1-\alpha. \tag{36}$$

In addition, for any $\delta \in (0,1)$ satisfying

$$e_n(\delta) \leq f_{\min}\big(\Delta - \overline{\beta}_n\big), \tag{37}$$

for every pilot realization in $\mathcal{G}_p$ and every realization of V, with probability at least $1-\delta$ over the calibration sample,

$$1-\alpha-e_n(\delta) \leq F(\tau) \leq 1-\alpha+e_n(\delta)+\frac{2(1+c)}{1-c}\,\widehat{\kappa}\kappa\,\frac{B_n}{\rho}. \tag{38}$$

Thus the final term in (38) is the rounding-induced contribution to the upper endpoint. The total upper excess also contains $e_n(\delta)$. Jointly over the pilot, shared offset, and calibration sample, (38) holds with probability at least $1-\delta_p-\delta$.

PROOF OF PROPOSITION 6

Put $r := 1-\alpha$ and let $\widehat{F}_1$ be the empirical distribution function of the first pilot half. Continuity of F implies that the first-half observations are distinct almost surely, and hence $\widehat{F}_1(\widehat{q}_p) = k_p/m$. Since $k_p = \lceil r(m+1)\rceil$ and $r \in (0,1)$,

$$0 < \frac{k_p}{m} - r < \frac{2}{m}. \tag{39}$$

Let $t_m := \sqrt{\log(6/\delta_p)/(2m)}$ and define $E_0 := \{\|\widehat{F}_1 - F\|_\infty \leq t_m\}$. The Dvoretzky–Kiefer–Wolfowitz inequality gives $\mathbb{P}(E_0) \geq 1-\delta_p/3$. On E_0,

$$\begin{aligned}|F(\widehat{q}_p)-r| &\leq \|\widehat{F}_1 - F\|_\infty + \left|\frac{k_p}{m} - r\right| \\ &< t_m + \frac{2}{m} = e_m(\delta_p/3) \leq \frac{f_{\min}\Delta}{2}.\end{aligned} \tag{40}$$

We claim that (40) implies $|\widehat{q}_p - q| \leq \Delta/2$. If $\widehat{q}_p > q+\Delta/2$, then monotonicity and the lower density bound on I give

$$F(\widehat{q}_p) \geq F(q+\Delta/2) = F(q) + \int_q^{q+\Delta/2} f(t)\,dt \geq r + \frac{f_{\min}\Delta}{2},$$

contradicting (40). The case $\widehat{q}_p < q-\Delta/2$ is symmetric. Because $h \leq \Delta/2$, it follows that, on E_0, $W = [\widehat{q}_p - h, \widehat{q}_p + h] \subseteq [q-\Delta, q+\Delta] = I$. Therefore, conditionally on the first pilot half,

$$p_W := \mathbb{P}(S \in W \mid \widehat{q}_p) = F(\widehat{q}_p + h) - F(\widehat{q}_p - h) = \int_{\widehat{q}_p - h}^{\widehat{q}_p + h} f(t)\,dt$$

satisfies

$$2hf_{\min} \leq p_W \leq 2hf_{\max}. \tag{41}$$

The two pilot halves are independent, so, conditionally on the first half, $N_W \sim \mathrm{Bin}(m, p_W)$. For $c \in (0,1)$, the multiplicative Chernoff bounds yield

$$\begin{aligned}\mathbb{P}[N_W \geq (1+c)mp_W \,|\, \widehat{q}_p] &\leq \exp\left(-\frac{c^2 mp_W}{3}\right), \\ \mathbb{P}[N_W \leq (1-c)mp_W \,|\, \widehat{q}_p] &\leq \exp\left(-\frac{c^2 mp_W}{2}\right).\end{aligned}$$

On E_0, (41) and (30) imply $mp_W \geq 2mhf_{\min} \geq 3\log(3/\delta_p)/c^2$. Consequently, after conditioning on the first half, each of the two tail probabilities on E_0 is at most $\delta_p/3$. Thus the event

$$\mathcal{G}_p := E_0 \cap \{(1-c)mp_W < N_W < (1+c)mp_W\}$$

has probability at least $1-\delta_p$. On $\mathcal{G}_p$, $N_W > (1-c)mp_W > 0$. Since N_W is an integer, $N_W \geq 1$, and the safeguard $\max\{N_W, 1\}$ is inactive. Dividing the preceding strict inequalities by $(1+c)2hm$ and applying (41) gives

$$\frac{1-c}{1+c} f_{\min} \leq \widehat{f}_{\mathrm{L}} \leq f_{\max}.$$

This proves (31). The only role of the safeguard is to make the protocol well defined on $\mathcal{G}_p^c$.

Fix a pilot realization in $\mathcal{G}_p$. Since $\widehat{\kappa} \geq \kappa = f_{\max}/f_{\min}$,

$$\beta = \frac{2\widehat{\kappa}B_n}{\widehat{f}_{\mathrm{L}}\rho} \geq \frac{2\widehat{\kappa}B_n}{f_{\max}\rho} \geq \frac{2B_n}{f_{\min}\rho}.$$

The lower bound on $\widehat{f}_{\mathrm{L}}$ similarly gives $\beta \leq \frac{2(1+c)}{1-c}\widehat{\kappa}B_n/(f_{\min}\rho) = \overline{\beta}_n$. This proves (32).

Conditional on the public pilot, β is fixed, while $u = \beta V$ is uniform on $[0,\beta)$ and independent of the two calibration samples. Conditions (33) therefore permit application of Theorem 5(i). The probability in the following display is over the two calibration samples and the shared V:

$$\mathbb{P}_{\mathcal{D},\mathcal{D}',V}[\tau(\mathcal{D}) \neq \tau(\mathcal{D}') \,|\, \mathcal{D}_p] \leq \frac{B_n}{f_{\min}\beta} + 4\exp\left(-\frac{n f_{\min}^2 \Delta^2}{2}\right) \leq \frac{\rho}{2} + \frac{\rho}{2} = \rho.$$

Theorem 5(iii) then gives the corresponding prediction-set-map claim. On $\mathcal{G}_p^c$ the mismatch probability is at most one, so averaging over the pilot gives

$$\mathbb{P}(\tau \neq \tau') \leq \mathbb{P}(\mathcal{G}_p)\rho + \mathbb{P}(\mathcal{G}_p^c) = \rho + (1-\rho)\mathbb{P}(\mathcal{G}_p^c) \leq \rho + (1-\rho)\delta_p.$$

If ρ in the construction is replaced by $\rho_0 = (\rho_\star - \delta_p)/(1-\delta_p)$, the last expression is exactly $\rho_\star$.

For marginal validity, fix arbitrary realizations of $\mathcal{D}_p$ and V. Then (β, u) is fixed independently of the calibration sample and test observation, and upward rounding gives $\tau \geq \widetilde{\tau}_n$ deterministically. By exchangeability and continuity of the $n+1$ scores,

$$\mathbb{P}_{\mathcal{D},(X_{n+1},Y_{n+1})}[Y_{n+1} \in \mathcal{C}_\tau(X_{n+1}) \,|\, \mathcal{D}_p, V] \geq \mathbb{P}[S_{n+1} \leq S_{(k_n)}] = \frac{k_n}{n+1} \geq 1-\alpha.$$

This proves (36) for every pilot and shared-offset realization, including those outside $\mathcal{G}_p$.

Finally, fix a pilot realization in $\mathcal{G}_p$ and a realization of V. By (32), $\beta \leq \overline{\beta}_n$. Hence (37) implies

$$e_n(\delta) \leq f_{\min}(\Delta - \overline{\beta}_n) \leq f_{\min}(\Delta - \beta).$$

Theorem 5(ii) therefore gives, with probability at least $1-\delta$ over the calibration sample,

$$1 - \alpha - e_n(\delta) \leq F(\tau) \leq 1 - \alpha + e_n(\delta) + f_{\max}\beta.$$

Using $\beta \leq \overline{\beta}_n$ and $f_{\max}/f_{\min} = \kappa$,

$$f_{\max}\beta \leq f_{\max}\overline{\beta}_n = \frac{2(1+c)}{1-c}\,\widehat{\kappa}\kappa\,\frac{B_n}{\rho}.$$

This proves (38). A union bound over pilot and calibration failure gives the asserted joint probability $1 - \delta_p - \delta$. □

PROOF OF PROPOSITION 2 (PLUG-IN WIDTH)

Write $\widetilde{\tau}, \widetilde{\tau}'$ for the two order statistics, $\Delta = \widetilde{\tau} - \widetilde{\tau}'$, and $k = \lceil (1-\alpha)(n+1) \rceil$. Since f is continuous and positive at q, the central order statistic satisfies $\sqrt{n}(\widetilde{\tau} - q) \Rightarrow \mathcal{N}(0, \alpha(1-\alpha)/f(q)^2)$ (Van der Vaart, 2000, Cor. 21.5). By independence, $\sqrt{n}\,\Delta \Rightarrow \mathcal{N}(0, 2\alpha(1-\alpha)/f(q)^2)$.

(i) Conditionally on $(\widetilde{\tau}, \widetilde{\tau}', \beta_n)$, the shared offset $u \sim \mathrm{Unif}[0, \beta_n)$ places a grid point in $(\widetilde{\tau} \wedge \widetilde{\tau}', \widetilde{\tau} \vee \widetilde{\tau}']$ with probability exactly $\min(|\Delta|/\beta_n, 1)$, and the outputs differ iff this happens. Now $|\Delta|/\beta_n = \big(\sqrt{n}f(q)|\Delta|/\sqrt{2\alpha(1-\alpha)}\big) \cdot \sqrt{2\alpha(1-\alpha)}/\big(\sqrt{n}f(q)\beta_n\big)$. The first factor converges to $|Z|$, and $\sqrt{n}f(q)\beta_n = c_0\sqrt{n}B_n f(q)/(\hat{f}_n\rho) \to_p c_0\sqrt{2\alpha(1-\alpha)}/\rho$ since $\sqrt{n}B_n \to \sqrt{2\alpha(1-\alpha)}$ and $\hat{f}_n \to_p f(q)$. By Slutsky, $|\Delta|/\beta_n \Rightarrow \rho|Z|/c_0$. Since $\min(\cdot, 1)$ is bounded and continuous, $\mathbb{P}[\tau \neq$

$\tau'] = \mathbb{E}\min(|\Delta|/\beta_n, 1) \to \mathbb{E}\min(\rho|Z|/c_0, 1) \le (\rho/c_0)\mathbb{E}|Z| = \rho\sqrt{2/\pi}/c_0$. With $c_0 = \sqrt{2/\pi}$ the limit is $\mathbb{E}\min\big(\sqrt{\pi/2}\,\rho|Z|, 1\big) = \rho(1 + o(1))$ as $\rho \to 0$.

(ii) Given $(\widetilde{\tau}, \beta_n)$, the up-rounding excess $\tau - \widetilde{\tau} = \beta_n\lceil(\widetilde{\tau} - u)/\beta_n\rceil + u - \widetilde{\tau}$ has normalized version $U := (\tau - \widetilde{\tau})/\beta_n \sim \mathrm{Unif}[0, 1)$ in u, independent of $(\widetilde{\tau}, \beta_n)$ (the excess $U\beta_n$ itself is not independent of a random β_n). Fix $0 < \varepsilon' \le f(q)/2$ and $\eta > 0$ with $|f(t) - f(q)| \le \varepsilon'$ on $[q - \eta, q + \eta]$. On $A_n = \{|\widetilde{\tau} - q| \le \eta/2,\ \beta_n \le \eta/2\}$, $|F(\widetilde{\tau} + U\beta_n) - F(\widetilde{\tau}) - f(q)U\beta_n| \le \varepsilon' U\beta_n$. Also, $\mathbb{P}(A_n^c) \le 2e^{-2n(f(q)\eta/4 - 2/n)^2} = o(B_n)$ by Lemma 3 once $c_0 B_n/(f_0\rho) \le \eta/2$, which the deterministic bound $\beta_n \le c_0 B_n/(f_0\rho)$ gives for large n, and F-increments are at most 1, while $\mathbb{E}[1/\hat{f}_n] \to 1/f(q)$ by bounded convergence ($1/\hat{f}_n \le 1/f_0$). Hence $\mathbb{E}[F(\tau) - F(\widetilde{\tau})] = f(q)\mathbb{E}[U]\,\mathbb{E}[\beta_n](1 + O(\varepsilon'/f(q))) + o(B_n) = (1 + o(1))c_0 B_n/(2\rho)$ after letting $\varepsilon' \downarrow 0$, and $\mathbb{E}F(\widetilde{\tau}) = k/(n + 1)$ exactly.

(iii) On $A_n \cap \{\hat{f}_n \ge f(q) - \varepsilon'\}$, whose probability tends to one, $F(\tau) - F(\widetilde{\tau}) \le (f(q) + \varepsilon')\beta_n \le \frac{f(q)+\varepsilon'}{f(q)-\varepsilon'} \cdot \frac{c_0 B_n}{\rho} = c_0(1 + O(\varepsilon'/f(q)))B_n/\rho$, and $\mathbb{P}[F(\widetilde{\tau}) \le k/(n + 1) + e_n(\delta)] \ge 1 - \delta$ as in Theorem 5(ii). A union bound gives (iii). □

Proof of Proposition 3

Let $U^+ := F(\widetilde{\tau}^+)$, which is $\mathrm{Beta}(k^+, n+1-k^+)$ with mean $m^+ := k^+/(n+1) \in [p_+, p_+ + 1/(n+1))$. (i) The sample has no ties almost surely, so $\widehat{F}_n(\widetilde{\tau}^+) = k^+/n \ge p_+$, and the Dvoretzky–Kiefer–Wolfowitz inequality gives, with probability at least $1 - \delta$, $F(\widetilde{\tau}^+) \ge \widehat{F}_n(\widetilde{\tau}^+) - \sqrt{\log(2/\delta)/(2n)} \ge p_+ - e_n(\delta) + 2/n > p$. Upward rounding gives $F(\tau^+) \ge F(\widetilde{\tau}^+)$ for every offset. (ii) Let E be the event that both raw cutoffs satisfy $|U_j^+ - p| < f_{\min}\Delta$. Since $|m^+ - p| \le e_n(\delta) + 1/(n + 1) \le 2e_n(\delta) \le f_{\min}\Delta/2$, this event contains $\{|U_j^+ - m^+| < f_{\min}\Delta/2,\ j = A, B\}$, whose complement has probability at most $4\exp\{-2n(f_{\min}\Delta/2 - 2/n)_+^2\}$ by the argument of Lemma 3 applied at rank k^+. On E both raw cutoffs lie in I, and the proof of Theorem 5(i) applies verbatim with U^+ in place of U: Lemma 2, then (4), then $\mathbb{E}|U^+ - U'^+| \le \sqrt{2\,\mathrm{Var}(U^+)} \le B_n(\alpha_+)$, where the variance bound (9) holds with α_+ in place of α because $m^+ \in [p_+, p_+ + 1/(n + 1))$. (iii) On the DKW event of (i), $|U^+ - p_+| \le e_n(\delta)$, so $|U^+ - p| \le 2e_n(\delta) \le f_{\min}\Delta/2$, which localizes $\widetilde{\tau}^+$ to within $\Delta/2$ of q by (4), and $[\widetilde{\tau}^+, \widetilde{\tau}^+ + \beta] \subseteq I$ because $\beta \le \Delta/2$. The upper density bound then gives $F(\tau^+) \le F(\widetilde{\tau}^+) + f_{\max}\beta \le p + 2e_n(\delta) + f_{\max}\beta$, and the lower bound is (i), and both hold for every offset because the DKW event depends only on the sample. For the consequence, substitute $\beta = 2B_n(\alpha_+)/(f_{\min}\rho)$ into (ii) to obtain mismatch at most $\rho/2 + \rho/2$, and into (iii) to obtain the cap $2e_n(\delta) + 2\kappa B_n(\alpha_+)/\rho$. Since $\alpha_+(1 - \alpha_+) \le \alpha(1 - \alpha) + e_n(\delta)$, $B_n(\alpha_+) = O(\sqrt{\alpha(1 - \alpha)/n} + \sqrt{e_n(\delta)/n})$, and the second term is $O(n^{-3/4}\log^{1/4}(1/\delta))$, which is of lower order. □

Proof of Proposition 4

(i) is the tower property followed by Markov's inequality. (ii) Since (V, β_n) is a function of the seed and the pilot, both independent of the calibration samples, conditionally on (V, β_n) the grid is fixed and the two raw thresholds are i.i.d. with their unconditional law, with $\sqrt{n}f(q)(\widetilde{\tau} - q) \Rightarrow \mathcal{N}(0, \alpha(1 - \alpha))$ (Van der Vaart, 2000, Cor. 21.5). The outputs coincide iff both raw thresholds fall in the same cell, and cells are intervals. Conditionally on (V, β_n) the grid is fixed, and by Pólya's theorem the distribution function of the normalized threshold converges to Φ uniformly, so the probability a_j of cell j under the law of $\widetilde{\tau}$ and its $\mathcal{N}(q, \sigma_{S,n}^2)$ mass $p_j(V)$ satisfy $\max_j |a_j - p_j(V)| \le 2d_n$ with $d_n \to 0$ the uniform distance, whence $|\sum_j a_j^2 - \sum_j p_j(V)^2| \le \max_j |a_j - p_j(V)| \sum_j (a_j + p_j(V)) \le 4d_n \to 0$. This gives $\mathbb{P}[\tau = \tau' \mid V, \beta_n] - \sum_j p_j(V)^2 \to 0$ uniformly in (V, β_n), even though the cells move with n. Because $\sqrt{n}f(q)\beta_n \to_p c_0\sqrt{2\alpha(1 - \alpha)}/\rho$, the cells other than the two adjacent to the boundary nearest q lie at distance at least $\beta_n/2$ from q and carry total mass at most $2\Phi(-c_0/(\sqrt{2}\rho))(1 + o(1))$. Dropping them leaves $1 - w^2 - (1 - w)^2 = 2w(1 - w)$ with $w = \Phi(-\ell(V)/\sigma_{S,n})$, up to that error. (iii) With a uniform shared phase, all M outputs coincide iff no grid point falls in $[\min_j \widetilde{\tau}_j, \max_j \widetilde{\tau}_j]$, which by the argument of Lemma 2 has probability exactly $1 - \mathbb{E}\min(R/\beta_n, 1)$ for the sample range R. The joint CLT gives $\sqrt{n}f(q)R/\sqrt{\alpha(1 - \alpha)} \Rightarrow R_M$, and

Slutsky yields the limit $1-\mathbb{E}\min\{\rho R_M/(c_0\sqrt{2}),1\}$. Writing $\mathbb{E}\min(aR_M,1) = a\,d_M - \mathbb{E}(aR_M - 1)_+$ with $a=\rho/(c_0\sqrt{2})$, and likewise for $M=2$, gives $1-(d_M/d_2)\pi_2$ up to the two truncation remainders, each of the form $\mathbb{E}[(aR-1)_+] \le \mathbb{E}[aR\,\mathbf{1}\{aR>1\}] \le a(\mathbb{E}R^2)^{1/2}\mathbb{P}(aR>1)^{1/2}$ with $\mathbb{P}(aR_M>1) \le \binom{M}{2}\mathbb{P}(a|Z_1-Z_2|>1)$ by a union bound, so both vanish as $\rho\sqrt{\log M}\to 0$. (iv) Under a location shift the two raw thresholds have the same limiting fluctuation and centers θ apart, so Proposition 2(i) applies with $\Delta+\theta$ in place of Δ, and $\min(|\Delta+\theta|,\beta_n) \le |\theta| + \min(|\Delta|,\beta_n)$. Without the location-shift assumption the two fluctuation scales differ and no bound in $|\theta|$ alone holds. Coverage of each analyst is the unchanged marginal statement of Theorem 5(ii). □

Proposition 7 (Pilot-centred grid in score units)**.** *Let a public pilot of n_p i.i.d. scores from the analysts' common law F, independent of both calibration samples, have conformal quantile $\hat{q}_p$. Let β be a width determined by the pilot, and let both analysts round upward onto the deterministic grid with boundaries $\hat{q}_p \pm \beta/2 + \beta\mathbb{Z}$ (no seed). Each analyst keeps marginal validity for every pilot. Conditionally on the pilot, both analysts deploy $\hat{q}_p+\beta/2$ whenever both raw cutoffs lie in $(\hat{q}_p-\beta/2,\ \hat{q}_p+\beta/2]$, so*

$$\mathbb{P}(\tau_A \ne \tau_B \mid \textit{pilot}) \le 2r_p - r_p^2, \qquad r_p := H_n\big(F(\hat{q}_p-\beta/2)\big) + 1 - H_n\big(F(\hat{q}_p+\beta/2)\big),$$

where r_p, the probability that one analyst's raw cutoff leaves the central cell, is exact given the pilot, and H_n is the $\mathrm{Beta}(k, n{+}1{-}k)$ distribution function. Under the conditions of Proposition 2 at the plug-in width ($c_0=1$), where $\beta_n/\sigma_{S,n}\to c := \sqrt{2}/\rho$, if $n_p/(\rho^2 n)\to\infty$ then the pilot's location error is negligible relative to the cell and the mismatch tends to the centred lattice value $1-\sum_{j\in\mathbb{Z}}\{\Phi((j+\frac{1}{2})c)-\Phi((j-\frac{1}{2})c)\}^2$. This value is positive at every fixed ρ (3.1×10^{-12} at $\rho{=}.1$). Identity tends to one if in addition $\rho=\rho_n\to 0$ with $\rho_n\sqrt{n/n_p}\to 0$. Under a common location shift θ of the analysts relative to the pilot, identity follows the same lattice law with its centre displaced by θ. It is near one while $|\theta|$ is small compared with $\beta/2-|\hat{q}_p-q|$, and it falls toward one half when a cell boundary reaches the analysts' quantile, with period β in θ. (Proof below.)

Proof of Proposition 7

Conditionally on the pilot the grid is fixed and independent of both calibration samples, so upward rounding keeps each analyst's marginal validity (Theorem 5(ii)). If both raw cutoffs lie in the central cell, both outputs equal $\hat{q}_p+\beta/2$. Given the pilot, the two raw cutoffs are independent, and by the probability integral transform each leaves the central cell with probability exactly r_p. Hence $\mathbb{P}(\tau_A=\tau_B \mid \text{pilot}) \ge (1-r_p)^2$, which is the display. For the limit, conditionally on the pilot, $\mathbb{P}(\tau_A=\tau_B \mid \text{pilot}) = \sum_j p_j^2$ over the cells of the grid, and by the order-statistic central limit theorem (Van der Vaart, 2000, Cor. 21.5) the cell masses are those of $\mathcal{N}(q,\sigma_{S,n}^2)$ up to $o(1)$. The pilot quantile satisfies $(\hat{q}_p-q)/\sigma_{S,n} = O_p(\sqrt{n/n_p})$, which is $o_p(\beta_n/\sigma_{S,n}) = o_p(c)$ when $\rho\sqrt{n/n_p}\to 0$. The grid is then asymptotically centred at q, and bounded convergence gives the stated limit. When also $c\to\infty$, the central cell carries all the mass in the limit. The shift statement follows by replacing q with $q+\theta$. □

Asymptotics of D_n. Let U,U' be i.i.d. $\mathrm{Beta}(k,n{+}1{-}k)$. Their variance is $m_n(1-m_n)/(n+2)$ with $m_n=k/(n+1)$, which is $\sigma_F^2(1+O(1/n))$, and $(U-m_n)/\sigma_F$ converges in law to a standard normal, since U is the k-th order statistic of n uniforms and k is within one rank of the sample p-quantile (Van der Vaart, 2000, Cor. 21.5). Hence $(U-U')/\sigma_F \Rightarrow \mathcal{N}(0,2)$, and because its second moments converge to 2, the family is uniformly integrable and $D_n/\sigma_F = \mathbb{E}|U-U'|/\sigma_F \to \mathbb{E}|\sqrt{2}Z| = 2/\sqrt{\pi}$. This is the expansion $D_n = (2/\sqrt{\pi})\sigma_F(1+o(1))$ of Section 2, and it gives $|U-U'|/D_n \Rightarrow \sqrt{\pi/2}\,|Z|$, which part (iii)(c) below uses.

Proof of Theorem 2

Throughout, G is right-continuous and nondecreasing and $Q^+(v) := Q_G^+(v) = \inf\{t: G(t)>v\}$ with $\inf\varnothing = +\infty$. Part (iii)(c) below is the shifted-pilot extension stated after the theorem in Section 4, and the case-(b) bound in (iv) is the pilot-error extension: in case (iii)(b) the excess coverage is at most $b + 2\omega + 2\lambda/n_p$ for every offset on the event $\|\widehat{F}_{n_p}-F\|_\infty \le \omega$, of probability at least $1-2\exp(-2n_p\omega^2)$.

(i) Let $v \geq G(t)$. Every $t' < t$ has $G(t') \leq G(t) \leq v$, so no point below t belongs to $\{t' : G(t') > v\}$, and $Q^+(v) \geq t$. Since $\mathrm{rd}_{u,b}(x) \geq x$, each analyst's deployed $\tau = Q^+(\mathrm{rd}_{u,b}(G(\widetilde{\tau})))$ satisfies $\tau \geq \widetilde{\tau}$ for every G and u. The sets are nested in the threshold, so $\mathbb{P}(Y_{n+1} \in \mathcal{C}_\tau(X_{n+1}) \mid G, u) \geq \mathbb{P}(S_{n+1} \leq S_{(k)}) = k/(n+1)$ by exchangeability, because (G, u) is independent of the calibration and test data. Equal rounded levels give equal thresholds, and hence equal maps.

(ii) The outputs are deterministic functions of the rounded levels, and the rounded levels differ exactly when a point of $u + b\mathbb{Z}$ lies between $G(\widetilde{\tau}_A)$ and $G(\widetilde{\tau}_B)$. By Lemma 2 this has probability $\min\{|G(\widetilde{\tau}_A) - G(\widetilde{\tau}_B)|/b, 1\}$ given the raw cutoffs and G.

(iii) (a) With $G = F$, the probability integral transform makes $F(\widetilde{\tau}_A), F(\widetilde{\tau}_B)$ independent $\mathrm{Beta}(k, n+1-k)$, so taking expectations in (ii) and using $\min\{x, 1\} \leq x$ gives D_n/b. (b) Fix $t \leq t'$. Since $\|G - \widehat{F}_{n_p}\|_\infty \leq \lambda/n_p$, we have $0 \leq G(t') - G(t) \leq \widehat{F}_{n_p}(t') - \widehat{F}_{n_p}(t) + 2\lambda/n_p$, and $\mathbb{E}[\widehat{F}_{n_p}(t') - \widehat{F}_{n_p}(t)] = F(t') - F(t)$ because $n_p\widehat{F}_{n_p}(t') - n_p\widehat{F}_{n_p}(t) \sim \mathrm{Bin}(n_p, F(t') - F(t))$. The pilot is independent of the calibration samples. Conditioning on the raw cutoffs and averaging over the pilot therefore gives $\mathbb{E}|G(\widetilde{\tau}_A) - G(\widetilde{\tau}_B)| \leq \mathbb{E}|F(\widetilde{\tau}_A) - F(\widetilde{\tau}_B)| + 2\lambda/n_p = D_n + 2\lambda/n_p$, and (ii) completes the proof. (c) The same computation with the pilot law F_{pil} gives $\mathbb{E}|F_{\mathrm{pil}}(\widetilde{\tau}_A) - F_{\mathrm{pil}}(\widetilde{\tau}_B)| + 2\lambda/n_p$. For the last claim it suffices to exhibit laws under which the mismatch itself is large. If G is continuous and strictly increasing on an interval containing both raw cutoffs, distinct rounded levels give distinct thresholds, so (ii) holds with equality. For $L > 1$, let F_{pil} have density Lf on a fixed neighbourhood of q of F-mass at most $1/(2L)$ (and be continuous elsewhere), let $n_p/n \to \infty$, so that $\|G - F_{\mathrm{pil}}\|_\infty = o_p(D_n)$ by the DKW inequality, and let $b = D_n/\rho$. Both cutoffs lie in the neighbourhood with probability tending to one, and there $|G(\widetilde{\tau}_A) - G(\widetilde{\tau}_B)| = L|F(\widetilde{\tau}_A) - F(\widetilde{\tau}_B)| + o_p(D_n)$. Since $|F(\widetilde{\tau}_A) - F(\widetilde{\tau}_B)|/D_n$ converges in law to $\sqrt{\pi/2}\,|Z|$ with Z standard normal, the mismatch tends to $\mathbb{E}\min\{L\rho\sqrt{\pi/2}\,|Z|, 1\}$, which increases to one as $L \to \infty$ while $D_n/b = \rho$ stays fixed.

(iv) Let $v := \mathrm{rd}_{u,b}(G(\widetilde{\tau})) \in [G(\widetilde{\tau}), G(\widetilde{\tau}) + b)$ and $t^* := Q^+(v)$. If $t^* < \infty$, then $G(t^*-) \leq v$, since every $t < t^*$ has $G(t) \leq v$. With $G = F$ continuous this gives $F(t^*) = F(t^*-) \leq v$, so $F(\tau) - F(\widetilde{\tau}) \leq v - F(\widetilde{\tau}) < b$. If $t^* = \infty$, then $F(\tau) = 1 \leq v$ and the same bound holds. Averaging over u, $v - F(\widetilde{\tau})$ is uniform on $[0, b)$, so the expected excess is at most $b/2$, with equality when no saturation at level 1 occurs. In case (b), on the event $\|\widehat{F}_{n_p} - F\|_\infty \leq \omega$, which by the DKW inequality (Massart, 1990) has probability at least $1 - 2e^{-2n_p\omega^2}$, continuity of F gives

$$F(t^*) = F(t^*-) \leq \widehat{F}_{n_p}(t^*-) + \omega \leq G(t^*-) + \tfrac{\lambda}{n_p} + \omega \leq v + \tfrac{\lambda}{n_p} + \omega,$$

while $F(\widetilde{\tau}) \geq \widehat{F}_{n_p}(\widetilde{\tau}) - \omega \geq G(\widetilde{\tau}) - \lambda/n_p - \omega \geq v - b - \lambda/n_p - \omega$. Subtracting gives $F(\tau) - F(\widetilde{\tau}) \leq b + 2\omega + 2\lambda/n_p$. If $t^* = \infty$, then $1 - \lambda/n_p \leq \sup G \leq v$ because $\widehat{F}_{n_p}$ reaches 1, so $F(\tau) = 1 \leq v + \lambda/n_p$ and the same bound holds. The event depends only on the pilot, so the bound holds for every offset and both analysts. □

Remark 2 (Coverage cost with a pilot map). In case (iii)(b), the case-(b) bound of (iv) gives $F(\tau) - F(\widetilde{\tau}) \leq b + 2\omega + 2\lambda/n_p$ for every offset on the pilot event $\|\widehat{F}_{n_p} - F\|_\infty \leq \omega$, and the DKW inequality for the analyst's own sample, on which $\widehat{F}_n(\widetilde{\tau}) = k/n$ almost surely, gives $F(\widetilde{\tau}) \leq k/n + \sqrt{\log(2/\delta)/(2n)} \leq 1 - \alpha + 2/n + \sqrt{\log(2/\delta)/(2n)}$ with probability at least $1 - \delta$. With $\omega = \sqrt{\log(4/\rho)/(2n_p)}$, $\delta = \rho/2$ and $b = (D_n + 2\lambda/n_p)/\rho$, each analyst's deployed coverage satisfies $F(\tau) \leq 1 - \alpha + (D_n + 2\lambda/n_p)/\rho + O\big(\sqrt{\log(1/\rho)}\,(n^{-1/2} + n_p^{-1/2})\big)$ with probability at least $1 - \rho$ (for $\rho \leq 1/32$, so that $\log(4/\rho) \leq 2\log(1/\rho)$), which is the coverage cost entered for this rule in Table 1 (with $\lambda = \frac{1}{2}$), while marginal validity holds by part (i). The realized $F(\tau)$ can fall below $1 - \alpha$, as for standard split conformal.

PROOF OF THEOREM 3

(i) is Theorem 2(i), which uses no property of G or of the phase beyond their independence from the calibration and test data.

(ii) Write $Z_{(1)} \leq \cdots \leq Z_{(n_p)}$ for the pilot order statistics, with the conventions $Z_{(i)} := -\infty$ for $i \leq 0$ and $Z_{(i)} := +\infty$ for $i \geq n_p + 1$, which correspond to $V_{(i)} = 0$ and $V_{(i)} = 1$ in the statement. Suppose $\widetilde{\tau} \geq Z_{(i_a)}$. Then $\widehat{F}_{n_p}(\widetilde{\tau}) \geq i_a/n_p$ and $G(\widetilde{\tau}) \geq (i_a - \lambda)/n_p > p - b/2$, because $i_a > n_p(p - b/2) + \lambda$.

Suppose $\widetilde{\tau} < Z_{(i_c)}$. Then $\widehat{F}_{n_p}(\widetilde{\tau}) \le (i_c - 1)/n_p$ and $G(\widetilde{\tau}) \le (i_c - 1 + \lambda)/n_p \le p + b/2$, because $i_c - 1 \le n_p(p + b/2) - \lambda$. If $i_a \le 0$, the lower condition holds trivially since $G \ge 0 > p - b/2$. If $i_c \ge n_p + 1$, then $n_p(p + b/2) - \lambda \ge n_p$, so $p + b/2 \ge 1 + \lambda/n_p \ge 1 \ge G$ and the upper condition holds trivially. So both transformed cutoffs lie in the central cell $(p - b/2, p + b/2]$, and both analysts deploy $Q^+(p + b/2)$, whenever $Z_{(i_a)} \le \widetilde{\tau}_j < Z_{(i_c)}$ for $j = A, B$. When $p + b/2 \ge 1$ this common threshold is $+\infty$, since $G \le 1 \le p + b/2$ everywhere, and both analysts deploy the full label set. As F is nondecreasing, $F(Z_{(i_a)}) < F(\widetilde{\tau}_j) < F(Z_{(i_c)})$ implies this event. For continuous F, $V_a := F(Z_{(i_a)})$ and $V_c := F(Z_{(i_c)})$ are the i_a-th and i_c-th order statistics of n_p i.i.d. uniforms, and $U_j := F(\widetilde{\tau}_j)$ are i.i.d. $\mathrm{Beta}(k, n+1-k)$ and independent of the pilot. Hence $\mathbb{P}(\tau_A = \tau_B) \ge \mathbb{E}[\mathbb{P}(V_a < U_A < V_c,\, V_a < U_B < V_c \mid V_a, V_c)] = \mathbb{E}[(H_n(V_c) - H_n(V_a))_+^2]$.

(iii) Let $t_n := \sqrt{\log(8/\rho)/(2n)}$ and $\bar{t}_{n_p} := \sqrt{\log(4/\rho)/(2n_p)}$. By the DKW inequality, the event $\mathcal{E}$ on which both calibration empirical distribution functions are within t_n of F and the pilot's is within $\bar{t}_{n_p}$ has probability at least $1 - 2 \cdot 2e^{-2nt_n^2} - 2e^{-2n_p\bar{t}_{n_p}^2} = 1 - \rho$. Since $\widehat{F}_{n,j}(\widetilde{\tau}_j) = k/n$ and $\widehat{F}_{n_p}(Z_{(i)}) = i/n_p$ almost surely, on $\mathcal{E}$ we have $|U_j - k/n| \le t_n$, $V_a \le i_a/n_p + \bar{t}_{n_p}$ when $1 \le i_a \le n_p$, and $V_c \ge i_c/n_p - \bar{t}_{n_p}$ when $1 \le i_c \le n_p$. From $0 \le k/n - p < 2/n$, $i_a/n_p \le p - b/2 + (\lambda+1)/n_p$ and $i_c/n_p > p + b/2 - \lambda/n_p$, the width condition yields $V_a < U_j < V_c$ for both analysts, and when $i_a \le 0$ or $i_c \ge n_p + 1$ the corresponding inequality $0 < U_j$ or $U_j < 1$ holds almost surely (the cases $i_a \ge n_p + 1$ and $i_c \le 0$ cannot occur, because $b/2 \ge (\lambda + 1)/n_p$ gives $n_p(p - b/2) + \lambda < n_p$ and $n_p(p+b/2) - \lambda > 0$). So both deploy $\tau = Q^+(p+b/2)$. If $\tau < \infty$ we have $G(\tau-) \le p + b/2 \le G(\tau)$, and on $\mathcal{E}$,

$$p \le p + \tfrac{b}{2} - \tfrac{\lambda}{n_p} - \bar{t}_{n_p} \le F(\tau) \le p + \tfrac{b}{2} + \tfrac{\lambda}{n_p} + \bar{t}_{n_p} \le p + b,$$

where the outer inequalities use $b/2 \ge \bar{t}_{n_p} + \lambda/n_p$. If $\tau = +\infty$, then $\sup G \le p + b/2$, and since $\widehat{F}_{n_p}$ reaches 1 we have $\sup G \ge 1 - \lambda/n_p$, so $F(\tau) = 1 \le p + b/2 + \lambda/n_p \le p + b$ and the display holds as well.

(iv) Given the pilot, the grid is fixed and the transformed cutoffs $G(\widetilde{\tau}_A), G(\widetilde{\tau}_B)$ are independent and identically distributed. The rounded levels are equal when both levels fall in the same cell, with all levels at or above 1 mapped to $+\infty$, which has probability $\sum_m \pi_m(G)^2$. Equal rounded levels give equal thresholds and hence identical maps, so this probability lower-bounds the probability of identical maps, and when G is continuous and strictly increasing on the range of the cutoffs, distinct rounded levels below 1 give distinct thresholds, so it equals the threshold identity. □

The certificate of Theorem 3(ii) is evaluated by two-dimensional Gauss–Legendre quadrature, writing $V_c = V_a + (1 - V_a)W$ with $W \sim \mathrm{Beta}(i_c - i_a,\, n_p + 1 - i_c)$ independent of $V_a \sim \mathrm{Beta}(i_a,\, n_p + 1 - i_a)$. The value is stable to 10^{-15} as the number of nodes varies and agrees with Monte Carlo. The smallest certified width for a target ρ is found by bisection over a lattice of step $1/(4n_p)$, since the certificate is nondecreasing in b (`certificates.py`). At $(n, n_p, \alpha) = (15{,}000, 2{,}000, .1)$: $D_n = .0027635$, $\Gamma(b_0) = .976$ at the implemented width $b_0 := \sqrt{2}\sigma_F/\rho = .0346$ of Appendix F ($\rho = .1$), and the certified widths at $\rho = .1, .05, .01$ are $.02650, .03063, .03963$, with certificates $.9025, .9522, .9911$.

Proof of Proposition 1

For each analyst $i \in \{A, B\}$, write $U_i := F(\widetilde{\tau}_i)$ for the realized coverage of its raw cutoff, which is $\mathrm{Beta}(k, n+1-k)$ by Section 2. U_A, U_B and the pilot are independent.

(i) Each analyst's $\tau = \max\{\widetilde{\tau}, T\} \ge \widetilde{\tau}$ and the sets are nested in the threshold, so conditionally on T, which is independent of the analyst's calibration and test points for every pilot law, $\mathbb{P}(Y_{n+1} \in \mathcal{C}_\tau(X_{n+1}) \mid T) \ge \mathbb{P}(S_{n+1} \le S_{(k)}) = k/(n+1) \ge 1 - \alpha$.

(ii) Both analysts deploy T exactly when $\widetilde{\tau}_A \le T$ and $\widetilde{\tau}_B \le T$. If exactly one raw cutoff exceeds T, one deployed threshold is T and the other is larger. If both exceed T, the deployed thresholds are the raw cutoffs, which are equal with probability zero by Proposition 5. Hence $\mathbb{P}(\tau_A = \tau_B) = \mathbb{P}(\widetilde{\tau}_A \le T, \widetilde{\tau}_B \le T)$. Since F is nondecreasing and continuous, $\{\widetilde{\tau}_i \le T\} \subseteq \{U_i \le F(T)\}$, and the difference of the two events is contained in $\{U_i = F(T)\}$, which is null because U_i has a continuous law and is independent of T. Conditionally on T the two events $\{U_i \le F(T)\}$ are independent with probability $H_n(F(T))$ each, so $\mathbb{P}(\tau_A = \tau_B) = \mathbb{E}[H_n(F(T))^2]$. For a pilot i.i.d. from F, $F(T) = F(Z_{(j)})$ is the j-th order statistic of n_p i.i.d. uniforms, $V_j \sim \mathrm{Beta}(j, n_p + 1 - j)$, which gives

the formula. For the mean coverage, $F(\tau_i) = F(\max\{\widetilde{\tau}_i, T\}) = \max\{U_i, F(T)\} = \max\{U_i, V_j\}$ with U_i independent of V_j.

(iii) Let $\mathcal{E}$ be the event that both calibration empirical distribution functions are within t_n of F and the pilot's is within $\bar{t}_{n_p}$. By the DKW inequality (Massart, 1990), $\mathbb{P}(\mathcal{E}^c) \le 2 \cdot 2e^{-2nt_n^2} + 2e^{-2n_p\bar{t}_{n_p}^2} = \rho/2 + \rho/2 = \rho$. On $\mathcal{E}$, since $\widehat{F}_{n,i}(\widetilde{\tau}_i) = k/n$ and $\widehat{F}_{n_p}(T) = j/n_p$ almost surely, $U_i \le k/n + t_n$ and $F(T) \ge j/n_p - \bar{t}_{n_p} \ge k/n + t_n$, so $U_i \le F(T)$ for both analysts, and by the null-event argument of (ii) both deploy T outside a null set. The same inequalities give $F(T) \ge k/n + t_n \ge 1 - \alpha$ and, using $j/n_p < k/n + t_n + \bar{t}_{n_p} + 1/n_p$ and $k/n < 1 - \alpha + 2/n$, $F(T) \le j/n_p + \bar{t}_{n_p} < 1 - \alpha + t_n + 2\bar{t}_{n_p} + 2/n + 1/n_p$.

(iv) For a pilot from any law, conditionally on T the argument of (ii) applies verbatim, because only the independence of U_A, U_B from T and the continuity of F were used, and gives identity $H_n(F(T))^2$. Since $\mathrm{Beta}(k, n{+}1{-}k)$ is the law of the k-th order statistic of n uniforms, $H_n(x) = \mathbb{P}(\mathrm{Bin}(n, x) \ge k)$. With $x = k/(n+1) - t$ and $t \ge 0$, $H_n(x) \le \mathbb{P}(\mathrm{Bin}(n, x)/n - x \ge k/n - k/(n+1) + t) \le \mathbb{P}(\mathrm{Bin}(n, x)/n - x \ge t) \le e^{-2nt^2}$ by Hoeffding's inequality, which equals ρ at $t = \sqrt{\log(1/\rho)/(2n)}$, so the identity is at most ρ^2 once $F(T) \le k/(n+1) - \sqrt{\log(1/\rho)/(2n)}$. □

The identity $\mathbb{E}[H_n(V_j)^2]$ and the mean coverage $\mathbb{E}\max\{U, V_j\} = 1 - \int_0^1 H_n(t)\,\mathbb{P}(V_j \le t)\,dt$ are evaluated by Gauss–Legendre quadrature against the $\mathrm{Beta}(j, n_p{+}1{-}j)$ density on an interval of fourteen standard deviations around its mean and by adaptive quadrature, and the smallest rank with identity at least $1 - \rho$ is found by bisection, since the identity is nondecreasing in j (`certificates.py`). At $(n, n_p, \alpha) = (15{,}000, 2{,}000, .1)$ the ranks for $\rho = .1, .05, .01$ are 1,822, 1,826 and 1,835, with identities .9122, .9507 and .9900 and mean coverages .9107, .9126 and .9171, and a Monte Carlo evaluation of $\mathbb{E}[H_n(V_j)^2]$ with 10^6 draws of V_j gives $.9124 \pm .0002$ at $j = 1{,}822$.

Remark 3 (Several analysts). Let $M \ge 2$ analysts share the same pilot and the same public inputs, each with its own independent calibration sample. Conditionally on the pilot, the realized coverages $U_1, \ldots, U_M$ of the raw cutoffs are i.i.d. $\mathrm{Beta}(k, n{+}1{-}k)$, so the proofs above go through with the square replaced by the M-th power: the centred grid has $\mathbb{P}(\tau_1 = \cdots = \tau_M) \ge \mathbb{E}[(H_n(V_c) - H_n(V_a))_+^M]$, and the pilot floor has $\mathbb{P}(\tau_1 = \cdots = \tau_M) = \mathbb{E}[H_n(V_j)^M]$ exactly, since all M analysts deploy T if and only if every raw cutoff is at most T, and any other pattern produces at least one raw cutoff, which coincides with another threshold with probability zero. Each analyst's validity is unchanged, because parts (i) hold analyst by analyst. For seeded rounding, the pairwise mismatch bound applied to analyst 1 against each of the others gives $\mathbb{P}(\tau_1 = \cdots = \tau_M) \ge 1 - (M-1)\rho$. Table S8 checks the three formulas for $M \in \{2, 5, 10\}$: at $(15{,}000, 2{,}000, .1)$ and $\rho = .1$ the floor's exact identity is .912/.878/.851 and the grid's certificate .903/.858/.822, against measured .907/.878/.847 and .944/.893/.851.

D Proof of Theorem 4

Scope of the scalar formulation. The theorem concerns exact replication of the scalar output index. It applies to a fixed, sample-independent sublevel family $\mathcal{C}_t(x) = \{y : s(x, y) \le t\}$ when the calibrator outputs the canonical index t, replicability is measured by equality of that index, and the admissible induced score laws include the hard family used below. For $S = s(X, Y)$,

$$\mathbb{P}\{Y \in \mathcal{C}_t(X)\} = \mathbb{P}\{S \le t\} = F_P(t).$$

Equal indices imply equal set maps, but the converse need not hold. Thus the theorem does not, without an additional injectivity or canonicalization assumption, give a lower bound when replicability is defined only by equality of set maps. Nor is an arbitrary sample-dependent or refitted nested family automatically covered.

Proof mechanism. The proof embeds a binomial statistic into the calibration problem, in the manner of the replicable-coin lower bounds of Impagliazzo et al. (2022, Theorem 7.1) and Hopkins et al. (2024, Theorem 3.7). It directly bounds the derivative of the seed-conditional acceptance probability and integrates the replicability constraint across the hard family, which yields the explicit constants and the factor $\alpha(1 - \alpha)$. Only the laws P_m with $m \in M$ are used, so the argument applies to every class of laws that contains this family, which is the form in which Theorem 4 is stated.

Proof. Put

$$p := 1 - \alpha, \qquad m_- := p - 2\varepsilon, \qquad m_+ := p + 2\varepsilon, \qquad M := [m_-, m_+],$$

and $\gamma := m_+ - m_- = 4\varepsilon$. Since $2\varepsilon \le \frac{1}{2}\min\{\alpha, p\}$, we have $m_- \ge p/2 > 0$ and $m_+ \le 1 - \alpha/2 < 1$, so $M \subset (0, 1)$.

Step 1: The hard family and its margin properties. For $m \in M$, let P_m have the following version of its Lebesgue density on $[0, 1]$:

$$f_m(s) = 2m\,\mathbf{1}\left\{0 \le s < \frac{1}{2}\right\} + 2(1-m)\,\mathbf{1}\left\{\frac{1}{2} \le s \le 1\right\}.$$

Its CDF is

$$F_m(t) = \begin{cases} 0, & t < 0, \\ 2mt, & 0 \le t \le \frac{1}{2}, \\ m + 2(1-m)(t - \frac{1}{2}), & \frac{1}{2} \le t \le 1, \\ 1, & t > 1. \end{cases}$$

Thus F_m is continuous and strictly increasing on $[0, 1]$. In particular, P_m is atomless even though its density may jump at $1/2$. The p-quantile is

$$q_m = \begin{cases} \dfrac{p}{2m}, & m \ge p, \\ \dfrac{1}{2} + \dfrac{p - m}{2(1-m)}, & m \le p. \end{cases}$$

If $m \ge p$, then $m \le p + 2\varepsilon \le \frac{3}{2}p$, so $q_m \ge 1/3$. If $m \le p$, write $d = p - m$. Since $0 \le d \le 2\varepsilon \le \alpha/2$, we have

$$q_m = \frac{1}{2} + \frac{d}{2(\alpha + d)} \le \frac{2}{3}.$$

The opposite bounds are immediate from the two cases. Hence $q_m \in [\frac{1}{3}, \frac{2}{3}]$ for every $m \in M$, and therefore $[q_m - \frac{1}{4}, q_m + \frac{1}{4}] \subset [0, 1]$.

In addition, $2m \ge 2m_- \ge p$ and $2(1-m) \ge 2(1-m_+) \ge \alpha$, and both possible density values are at most 2. Consequently, the displayed version of f_m lies in $[\min\{\alpha, p\}, 2]$ on all of $[0, 1]$, hence throughout $[q_m - \frac{1}{3}, q_m + \frac{1}{3}] \subset [0, 1]$ and throughout $[q_m - \frac{1}{4}, q_m + \frac{1}{4}]$. This proves the common margin claim for the entire family.

At the upper endpoint,

$$q_{m_+} = \frac{p}{2m_+} < \frac{1}{2}, \qquad \frac{1}{2} - q_{m_+} = \frac{\varepsilon}{m_+} > \varepsilon.$$

At the lower endpoint,

$$q_{m_-} = \frac{1}{2} + \frac{\varepsilon}{1 - m_-}, \qquad q_{m_-} - \frac{1}{2} = \frac{\varepsilon}{1 - m_-} > \varepsilon.$$

Because $\varepsilon \le 1/8$ and both endpoint quantiles belong to $[1/3, 2/3]$,

$$[q_{m_+} - \varepsilon, q_{m_+} + \varepsilon] \subset (0, \tfrac{1}{2}), \qquad [q_{m_-} - \varepsilon, q_{m_-} + \varepsilon] \subset (\tfrac{1}{2}, 1).$$

Consequently, P_{m_+} satisfies Assumption 1 with $\Delta = \varepsilon$, $f_{\min} = f_{\max} = 2m_+$, while P_{m_-} satisfies it with $\Delta = \varepsilon$, $f_{\min} = f_{\max} = 2(1 - m_-)$. Thus each endpoint separately has $\kappa = 1$, and the two constant density values both belong to $[2\min\{\alpha, 1 - \alpha\}, 2]$.

Step 2: Accuracy forces a binary decision. Define

$$B(\mathcal{D}; r) := \mathbf{1}\left\{\mathcal{A}(\mathcal{D}; r) < \frac{1}{2}\right\}.$$

Under P_{m_+}, every $t \ge 1/2$ satisfies

$$F_{m_+}(t) \ge m_+ = p + 2\varepsilon > p + \varepsilon.$$

Thus an accurate output must be below $1/2$, and hence

$$\mathbb{P}_{\substack{r\sim\nu \\ \mathcal{D}\sim P_{m_+}^{\otimes n}}}[B(\mathcal{D};r)=1] \geq 1-\delta.$$

Under P_{m_-}, every $t < 1/2$ satisfies

$$F_{m_-}(t) \leq m_- = p - 2\varepsilon < p - \varepsilon.$$

Thus an accurate output must be at least $1/2$, and hence

$$\mathbb{P}_{\substack{r\sim\nu \\ \mathcal{D}\sim P_{m_-}^{\otimes n}}}[B(\mathcal{D};r)=1] \leq \delta.$$

Finally,

$$\{B(\mathcal{D};r) \neq B(\mathcal{D}';r)\} \subseteq \{\mathcal{A}(\mathcal{D};r) \neq \mathcal{A}(\mathcal{D}';r)\}.$$

Therefore B is also ρ-replicable at every P_m, $m \in M$.

Step 3: Binomial reduction and derivative bound. Let

$$N := \sum_{i=1}^{n} \mathbf{1}\left\{S_i < \frac{1}{2}\right\}.$$

Under $P_m^{\otimes n}$, $N \sim \mathrm{Bin}(n,m)$, and the joint density of the sample, with respect to Lebesgue measure, is $2^n m^N (1-m)^{n-N}$ almost everywhere.

For $k \in \{0,\dots,n\}$, let Q_k be the law obtained by first choosing a k-subset J of $\{1,\dots,n\}$ uniformly and then, independently, drawing

$$S_i \sim \mathrm{Unif}[0,\tfrac{1}{2}) \quad (i \in J), \qquad S_i \sim \mathrm{Unif}[\tfrac{1}{2},1] \quad (i \notin J).$$

For every $m \in M$, the conditional law of $\mathcal{D}$ given $N = k$ is Q_k, which does not depend on m.

For fixed r, define

$$\psi_r(k) := \int_{[0,1]^n} \mathbf{1}\{B(\mathbf{s};r)=1\}\; Q_k(\mathrm{d}\mathbf{s}) \in [0,1]$$

and

$$h_r(m) := \mathbb{P}_{\mathcal{D}\sim P_m^{\otimes n}}[B(\mathcal{D};r)=1].$$

Conditioning on N gives

$$h_r(m) = \sum_{k=0}^{n} \psi_r(k) \binom{n}{k} m^k (1-m)^{n-k}.$$

Joint measurability of $\mathcal{A}$ implies that $r \mapsto \psi_r(k)$ is measurable. Hence $(r,m) \mapsto h_r(m)$ is jointly measurable. The endpoint conclusions give

$$\mathbb{E}_r h_r(m_-) \leq \delta, \qquad \mathbb{E}_r h_r(m_+) \geq 1-\delta.$$

For each fixed r, h_r is a polynomial. In what follows, $\mathbb{E}_m$, Cov_m, and Var_m refer to $N \sim \mathrm{Bin}(n,m)$. Differentiating the binomial probabilities gives

$$h_r'(m) = \mathbb{E}_m\left[\psi_r(N)\frac{N-nm}{m(1-m)}\right] = \frac{\mathrm{Cov}_m(\psi_r(N),N)}{m(1-m)}.$$

We have $\mathrm{Var}_m(N) = nm(1-m)$, and, since $0 \leq \psi_r(N) \leq 1$, $\mathrm{Var}_m(\psi_r(N)) \leq h_r(m)(1-h_r(m))$. Consequently, by Cauchy–Schwarz,

$$|h_r'(m)| \leq \frac{\sqrt{\mathrm{Var}_m(\psi_r(N))\,\mathrm{Var}_m(N)}}{m(1-m)} \leq \sqrt{\frac{n\,h_r(m)(1-h_r(m))}{m(1-m)}}. \tag{42}$$

Step 4: Good seeds and the transition. Fix $\eta \in (2\delta, \frac{1}{2})$ and let

$$G := \{r : \ h_r(m_-) \le \eta \text{ and } h_r(m_+) \ge 1-\eta\}.$$

Markov's inequality applied to $\mathbb{E}_r h_r(m_-) \le \delta$ and $\mathbb{E}_r[1-h_r(m_+)] \le \delta$ gives

$$\mathbb{P}_r(G) \ge 1 - 2\delta/\eta. \tag{43}$$

Set $v_* := \inf_{m\in M} m(1-m)$. Fix $r \in G$. By continuity and the intermediate value theorem the level sets below are nonempty and closed, so the extrema exist:

$$b_r := \min\{m \in M : h_r(m) = 1-\eta\}, \qquad a_r := \max\{m \in [m_-, b_r] : h_r(m) = \eta\},$$

and minimality of b_r, maximality of a_r, and continuity give $h_r(m) \in [\eta, 1-\eta]$ on $[a_r, b_r]$ (a value above $1-\eta$ before b_r, or below η after a_r, would create an earlier or later crossing). No monotonicity is assumed.

Step 5: Integrating replicability against the Fisher bound. Conditional on a fixed seed r, the two decisions based on independent samples are independent Bernoulli variables with success probability $h_r(m)$, so their disagreement probability is $2h_r(m)(1-h_r(m))$, and decision disagreement implies threshold disagreement. Hence $\mathbb{E}_r[2h_r(m)(1-h_r(m))] \le \rho$ for every $m \in M$, and by Tonelli the jointly measurable function $I(r) := \int_M 2h_r(m)(1-h_r(m))\,dm$ satisfies

$$\mathbb{E}_r I(r) \le \rho\,|M| = \rho\gamma. \tag{44}$$

For $r \in G$, (42) gives $|h_r'(m)| \le \sqrt{n\,h_r(m)(1-h_r(m))/v_*}$ on M, so with $\Psi(u) := \int_0^u 2\sqrt{t(1-t)}\,dt$ and the fundamental theorem of calculus for the polynomial h_r,

$$\begin{aligned} I(r) \ &\ge \ \int_{a_r}^{b_r} 2h_r(1-h_r)\,dm \ \ge \ \int_{a_r}^{b_r} \frac{2h_r(1-h_r)\,|h_r'|}{\sqrt{n\,h_r(1-h_r)/v_*}}\,dm \\ &= \ \sqrt{\frac{v_*}{n}} \int_{a_r}^{b_r} 2\sqrt{h_r(1-h_r)}\,|h_r'|\,dm \ \ge \ \sqrt{\frac{v_*}{n}}\,\big|\Psi(h_r(b_r)) - \Psi(h_r(a_r))\big| \ = \ \sqrt{\frac{v_*}{n}}\,J(\eta), \end{aligned}$$

where $J(\eta) = \Psi(1-\eta) - \Psi(\eta) = \int_\eta^{1-\eta} 2\sqrt{u(1-u)}\,du = (1-2\eta)\sqrt{\eta(1-\eta)} + \frac{1}{2}\arcsin(1-2\eta)$. Combining with (43) and (44), without requiring measurability of a_r, b_r,

$$\rho\gamma \ \ge \ (1-2\delta/\eta)\,J(\eta)\sqrt{v_*/n}, \qquad \text{so} \qquad n \ \ge \ \frac{J(\eta)^2(1-2\delta/\eta)^2\,v_*}{\rho^2\gamma^2}.$$

Step 6: Constants. For each $\theta \in (0,1]$, repeat the argument with $m_\pm^{(\theta)} := p \pm (1+\theta)\varepsilon$ and $\gamma_\theta = 2(1+\theta)\varepsilon$. The endpoint decisions of Step 2 remain forced because $F_{m_\pm^{(\theta)}}(\frac{1}{2}) = m_\pm^{(\theta)} \notin [p-\varepsilon, p+\varepsilon]$, and Steps 3–5 use only the binomial structure and the common margin bounds of Step 1, which hold for every $m \in [p-2\varepsilon, p+2\varepsilon]$. The endpoint-specific windows of Step 1 are not used. For every $m \in M \subseteq [p-2\varepsilon, p+2\varepsilon]$ and $2\varepsilon \le \frac{1}{2}\min\{\alpha, p\}$, we have $m \ge p/2$ and $1-m \ge \alpha$ when $m \le p$, and $1-m \ge \alpha/2$ with $m \ge p$ when $m \ge p$, so $v_* \ge \alpha(1-\alpha)/2$. The conclusion $n \ge J(\eta)^2(1-2\delta/\eta)^2 v_*/(\rho^2\gamma_\theta^2)$ holds for every $\theta > 0$, and letting $\theta \downarrow 0$ gives $\gamma = 2\varepsilon$. Hence

$$n \ \ge \ \frac{J(\eta)^2(1-2\delta/\eta)^2}{8}\,\frac{\alpha(1-\alpha)}{\varepsilon^2\rho^2}.$$

With $\eta = \frac{1}{8}$ and $\delta \le \frac{1}{32}$, $J(\frac{1}{8}) = \frac{3}{4}\sqrt{7/64} + \frac{1}{2}\arcsin\frac{3}{4} = .672$ and $(1-16\delta)^2 \ge \frac{1}{4}$, which gives the constant $.672^2/32 = .0141 \ge 1/72$. As $\delta \to 0$, choosing $\eta \to 0$ with $\delta/\eta \to 0$ gives $J(\eta) \to \pi/4$ and the constant tends to $\pi^2/128$. □

Step 7: Limitation of the hard-family reduction. The preceding hard family alone cannot yield the same lower bound for outputs that may be arbitrary measurable subsets of the score space, without the restriction to sublevel sets from the fixed nested family. The fixed set

$$A^\star = \left[0, \frac{1-\alpha}{2}\right] \cup \left[\frac{1}{2}, \frac{1}{2} + \frac{1-\alpha}{2}\right]$$

satisfies

$$P_m(A^\star) = m(1-\alpha) + (1-m)(1-\alpha) = 1-\alpha$$

for every $m \in M$. Thus a deterministic, perfectly replicable procedure can attain exact coverage on this particular family if it is allowed such outputs. This observation concerns the scope of this reduction. It is not a universal positive result for unrestricted set outputs. There is no vanishing-margin-width issue in the construction. The whole family has the common parameters established in Step 1. The convenient displayed choices correspond to $\kappa_{\text{common}} = 2/\min\{\alpha, 1-\alpha\}$. At $m = 1-\alpha$, the sharp ratio between the two density values is $\max\{\alpha, 1-\alpha\}/\min\{\alpha, 1-\alpha\}$. Hence the proof establishes the stated $\alpha(1-\alpha)/(\varepsilon^2\rho^2)$ lower bound, but does not by itself show that the κ^2 factor in the upper bound is minimax necessary.

PROOF OF COROLLARY 1 (SET-SIZE COST)

By a union bound, for every P in the class, $\mathbb{P}[F_P(\mathcal{A}(\mathcal{D}; r)) \in [1-\alpha,\ 1-\alpha+e]] \geq 1-2\delta$. Set $\alpha' := \alpha - e/2$ and $\varepsilon' := e/2$, so the band is $[1-\alpha'-\varepsilon',\ 1-\alpha'+\varepsilon']$ and $\mathcal{A}$ satisfies the hypotheses of Theorem 4 at $(\alpha', \varepsilon', \rho, \delta'{=}2\delta)$ over the class, which contains $\mathcal{P}_2(\alpha', \varepsilon') = \mathcal{P}_2(\alpha - e/2, e/2)$, once the side conditions hold. They hold: $\delta' = 2\delta < \frac{1}{8}$, and since $e \leq \min(\alpha, 1-\alpha)/4$ we have $\min(\alpha', 1-\alpha') \geq \min(\alpha, 1-\alpha) - e/2 \geq \frac{7}{8}\min(\alpha, 1-\alpha) \geq \frac{7}{2}e > 4\varepsilon'$, so $\varepsilon' \leq \min(\alpha', 1-\alpha')/4$. Theorem 4 with $\eta = \frac{1}{4}$ (admissible since $2\delta' = 4\delta < \eta$) gives $n \geq \frac{J(1/4)^2(1-16\delta)^2}{8}\,\alpha'(1-\alpha')/(\varepsilon'^2\rho^2)$, i.e. $e = 2\varepsilon' \geq \frac{J(1/4)(1-16\delta)}{\sqrt{2}}\sqrt{\alpha'(1-\alpha')/n}/\rho$ with $J(1/4) = \frac{1}{2}\sqrt{3/16} + \frac{1}{2}\arcsin\frac{1}{2} = .478$. For the simplified form: if $\alpha \leq \frac{1}{2}$ then $\alpha' \geq \frac{7}{8}\alpha$ and $1-\alpha' \geq 1-\alpha$, so $\alpha'(1-\alpha') \geq \frac{7}{8}\alpha(1-\alpha)$. If $\alpha > \frac{1}{2}$, then α' moves toward $\frac{1}{2}$, so $\alpha'(1-\alpha') \geq \alpha(1-\alpha)$ directly once $\alpha' \geq \frac{1}{2}$, and otherwise $\alpha' \in (\frac{1}{2} - \frac{e}{2}, \frac{1}{2}]$ gives $\alpha'(1-\alpha') \geq \frac{1}{4} - \frac{e^2}{4} \geq \frac{7}{8}\cdot\frac{1}{4} \geq \frac{7}{8}\alpha(1-\alpha)$. With $\delta \leq \frac{1}{32}$, $\frac{J(1/4)(1-16\delta)}{\sqrt{2}}\sqrt{7/8} \geq \frac{.478}{2\sqrt{2}} \cdot 0.935 = .158 \geq \frac{3}{20}$. More generally, applying Theorem 4 with any $\eta \in (4\delta, \frac{1}{2})$ gives

$$e \geq \frac{J(\eta)(1-4\delta/\eta)}{\sqrt{2}}\,\frac{\sqrt{\alpha'(1-\alpha')/n}}{\rho}.$$

As $\delta \downarrow 0$, choosing $\eta \downarrow 0$ with $\delta/\eta \to 0$ makes the coefficient tend to $\pi/(4\sqrt{2}) \approx .555$ in these α' units. Uniformly over the admissible e the coefficient in α units is at least $\pi\sqrt{7}/16 \approx .519$, and for fixed α and $e \to 0$ the ratio $\alpha'(1-\alpha')/\{\alpha(1-\alpha)\}$ tends to one. □

Scope. The corollary inherits the scope of Theorem 4: threshold replicability over a class of laws containing the two-piece family, with a shared seed. The two-sided band is supplied by the validity hypothesis on one side and the inflation cap on the other. Neither alone suffices, and the oracle rule $\tau \equiv F_P^{-1}(1-\alpha)$ shows the worst-case quantifier cannot be dropped. *Cardinality.* For finite $\mathcal{Y}$ write $H(t) := \mathbb{E}_X|\mathcal{C}_t(X)|$ and let e_H be a cap on the excess mean cardinality $H(\tau) - H(q)$ that holds with the corollary's probability. For every $t \geq q$, $H(t) - H(q) = \mathbb{E}_X \sum_y \mathbf{1}\{q < s(X,y) \leq t\} \geq \mathbb{P}\{q < s(X,Y) \leq t\} = F(t) - F(q)$, because the true label is one of the candidate labels. Hence $e \leq e_H$ with no further assumption, and the lower bound on e transfers to e_H with $g_{\min} = 1$. If H is absolutely continuous with respect to F on an interval that contains q and every threshold that can occur, with $dH/dF \geq g_{\min}$ there, the same argument gives $H(\tau) - H(q) \geq g_{\min}\{F(\tau) - F(q)\}$ and $e \leq e_H/g_{\min}$, which is sharper when $g_{\min} > 1$. For $\tau < q$ the coverage cap is automatic. If the thresholds are confined to the interval only with probability $1-\zeta$, the cap's failure probability becomes $\delta + \zeta$. The argument does not bound realized or expected inflation from below: the corollary constrains an achievable cap and leaves the realized excess free.

PROOF OF COROLLARY 2

Seedless rules. Put $b := 2\{t_n + \bar{t}_{n_p} + (\lambda+1)/n_p + 2/n\}$. For a pilot of n_p i.i.d. draws from any continuous F, Theorem 3(iii) gives, with probability at least $1-\rho$, one common threshold τ for both analysts with $1-\alpha \leq F(\tau) \leq 1-\alpha+b$. The centred grid at this width is therefore ρ-replicable, satisfies $\mathbb{P}\{F(\tau) \geq 1-\alpha\} \geq 1-\rho$, and has the coverage-inflation cap $e_{\text{pilot}} = b$ with probability at least $1-\rho$. Proposition 1(iii) gives the same three properties for the pilot floor at the rank stated there, whenever that rank is at most n_p, with the cap $t_n + 2\bar{t}_{n_p} + 2/n + 1/n_p$. For $\rho \leq 1/32$ we have $\log(8/\rho) \leq 2\log(1/\rho)$ and $\log(4/\rho) \leq 2\log(1/\rho)$, so $t_n \leq \sqrt{\log(1/\rho)/n}$ and

$\bar{t}_{n_p} \le \sqrt{\log(1/\rho)/n_p}$. Since also $\log(1/\rho) \ge 1$, $1/n \le n^{-1/2}$ and $1/n_p \le n_p^{-1/2}$, both caps are at most $2(\lambda+3)\sqrt{\log(1/\rho)}\,(n^{-1/2} + n_p^{-1/2})$.

Seed-only rules. Let a threshold calibrator whose shared randomness is independent of F meet the three requirements, with cap e, for every F in a class containing $\mathcal{P}_2(\alpha - e/2, e/2)$. If $e > \min(\alpha, 1-\alpha)/4$, the stated bound holds. Otherwise the calibrator satisfies the hypotheses of Corollary 1 with $\delta = \rho \le 1/32$, which gives $e \ge \frac{3}{20}\sqrt{\alpha(1-\alpha)/n}/\rho$.

Ratio. When the first term of the minimum is the smaller,

$$\frac{e_{\text{pilot}}}{e} \le \frac{2(\lambda+3)\sqrt{\log(1/\rho)}\,(n^{-1/2} + n_p^{-1/2})\,\rho}{\frac{3}{20}\sqrt{\alpha(1-\alpha)/n}} = \frac{40(\lambda+3)}{3\sqrt{\alpha(1-\alpha)}}\,\rho\sqrt{\log(1/\rho)}\,\big(1 + \sqrt{n/n_p}\big),$$

which tends to zero when $\rho\sqrt{\log(1/\rho)} \to 0$ and $n\rho^2 \log(1/\rho)/n_p \to 0$.

Pilot labels. The comparator of the last claim is a seed-only calibrator whose analysts each draw $n + n_p$ independent samples from F, as many labeled scores as the pilot rule sees in total. It is a calibrator on $n + n_p$ i.i.d. draws with shared randomness independent of F, so the argument above with $n + n_p$ in place of n gives the last bound. Against this comparator the ratio is at most $\frac{40(\lambda+3)}{3\sqrt{\alpha(1-\alpha)}}\,\rho\sqrt{\log(1/\rho)}\,(n^{-1/2} + n_p^{-1/2})\sqrt{n+n_p} \le \frac{40(\lambda+3)}{3\sqrt{\alpha(1-\alpha)}}\,\rho\sqrt{\log(1/\rho)}\,\big(2 + \sqrt{n/n_p} + \sqrt{n_p/n}\big)$, which for $n_p \le n$ is again $O(\rho\sqrt{\log(1/\rho)}\,(1 + \sqrt{n/n_p}))$. A common pilot given to a seed-only calibrator is shared data and is outside the scope of Theorem 4. □

Lemma 4 (Threshold and map identity)**.** *For a shared score s and thresholds τ, τ', equal thresholds give equal maps, so $I_\tau \le I$. If, whenever $\tau \ne \tau'$, the event $\{x : \exists y,\ \tau \wedge \tau' < s(x,y) \le \tau \vee \tau'\}$ has positive P_X-probability, which holds when the score has positive density on the interval between any two admissible thresholds, then distinct thresholds give distinct maps and $I_\tau = I$. In that case $\mathbb{P}[\mathcal{C}_\tau \not\equiv \mathcal{C}_{\tau'}] = \mathbb{P}[\tau \ne \tau']$, so lower bounds on the mismatch probability transfer to the deployed maps as well as upper bounds. If the hypothesis holds for every law of a class on the event that both thresholds lie in a set S, and $\mathbb{P}(\tau \notin \mathsf{S}) \le \zeta$ for every law of the class and each run, then map replicability at level ρ over the class implies threshold replicability at level $\rho + 2\zeta$. The two-piece family of Theorem 4 has density at least $\min\{\alpha, 1-\alpha\}$ on $[0,1]$, so the hypothesis holds there with $\mathsf{S} = [0,1]$. Under the two-sided accuracy of Theorem 4, a threshold whose coverage lies in $[1-\alpha-\varepsilon, 1-\alpha+\varepsilon]$ lies in $(0,1)$, so $\zeta = \delta$ and the theorem holds for the deployed maps with $\rho + 2\delta$ in place of ρ. Under the validity and cap hypotheses of Corollary 1, $F_P(\tau) \ge 1-\alpha$ excludes $\tau < 0$ and $F_P(\tau) \le 1 - \alpha + e < 1$ excludes $\tau \ge 1$, each with probability at least $1-\delta$, so $\zeta = 2\delta$ and the corollary holds for the deployed maps with $\rho + 4\delta$ in place of ρ.*

PROOF OF LEMMA 4

The map $\mathcal{C}_\tau$ is a function of τ, so equal thresholds give equal maps and $\{\mathcal{C}_\tau \not\equiv \mathcal{C}_{\tau'}\} \subseteq \{\tau \ne \tau'\}$. Let $\tau < \tau'$. On the event of the hypothesis some input x has a candidate label y with $s(x,y) \in (\tau, \tau']$, so $y \in \mathcal{C}_{\tau'}(x) \setminus \mathcal{C}_\tau(x)$ and the maps differ. If the score has positive density on $(\tau, \tau']$, then $\mathbb{P}\{\tau < s(X,Y) \le \tau'\} > 0$, so the true label is such a witness for a set of inputs of positive P_X-probability. Hence $\{\tau \ne \tau'\} = \{\mathcal{C}_\tau \not\equiv \mathcal{C}_{\tau'}\}$ up to a null event and $I_\tau = I$. For the second claim, $\{\tau \ne \tau'\} \subseteq \{\mathcal{C}_\tau \not\equiv \mathcal{C}_{\tau'}\} \cup \{\tau \notin \mathsf{S}\} \cup \{\tau' \notin \mathsf{S}\}$, so $\mathbb{P}[\tau \ne \tau'] \le \rho + 2\zeta$ for every law of the class. For the two-piece family, a threshold outside $[0,1]$ has realized coverage 0 or 1: the two-sided band of Theorem 4 excludes both values, with probability at least $1-\delta$ per run, and the one-sided validity of Corollary 1 excludes 0 while its cap excludes 1, each with probability at least $1-\delta$ per run, whence $\zeta = 2\delta$ there. □

E FULL CONFORMAL PREDICTION

Proposition 8 (Shared cutoff rounding does not stabilize full-conformal 1-NN)**.** *Consider full conformal prediction with the 1-nearest-neighbor score, calibration size $n \ge 2$, and data law P absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^2$. Let $c_{\mathcal{D}}(x,y)$ be the number of the n training scores that are at least the candidate's score after augmentation by (x,y) (the convention below), and for a cutoff m let $\widehat{\mathcal{C}}_{\mathcal{D}}^{(m)}(x) := \{y : c_{\mathcal{D}}(x,y) \ge m\}$. For every nonendpoint cutoff*

$m \in \{1, \dots, n\}$, two independent runs satisfy $\mathbb{P}[\widehat{\mathcal{C}}_{\mathcal{D}}^{(m)} \equiv \widehat{\mathcal{C}}_{\mathcal{D}'}^{(m)}] = 0$ (map equality read in the completed law). With any shared seed U and measurable cutoff $M(U)$, the agreement probability equals $\mathbb{P}[M(U) \in \{0, n+1\}]$ exactly. The usual rule $\{\widehat{p}_{\mathcal{D}} > \alpha\}$, where $\widehat{p}_{\mathcal{D}}$ is the full-conformal p-value, is the cutoff $\lfloor \alpha(n+1) \rfloor$, so agreement is zero for every $\alpha \in [1/(n+1), 1)$, with no integrality condition, and trivially one below. The condition $n \geq 2$ is sharp: for $n = 1$ the single training point and the candidate are each other's nearest neighbours, so $\widehat{p}_{\mathcal{D}} \equiv 1$ and the usual rule is the always-$\mathbb{R}$ map. The statement covers every sample-independent, seed-dependent nondecreasing transformation q_U of the p-value grid followed by a strict threshold a, since such a rule accepts exactly the candidates with $c_{\mathcal{D}}(x, y) \geq M(U) := \min\{c \in \{0, \dots, n\} : q_U((1+c)/(n+1)) > a\}$ (with $\min \varnothing := n+1$), a measurable cutoff when $u \mapsto q_u(v)$ is measurable for each grid value v. The conclusion concerns equality on all of $\mathbb{R}$. It does not cover data-dependent cutoff rules, agreement only on the support of P_X, or changes to the score-fitting procedure (stabilizing the refitted training map, as in replicable regression (Eaton et al., 2026), is outside its scope). (Full statement, conventions, and proof below.)

Full-conformal 1-NN convention. Let $\mathcal{D} = ((X_1, Y_1), \dots, (X_n, Y_n))$ and, for a candidate (x, y), write

$$(\widetilde{X}_i, \widetilde{Y}_i) = (X_i, Y_i), \quad 1 \leq i \leq n, \qquad (\widetilde{X}_{n+1}, \widetilde{Y}_{n+1}) = (x, y).$$

For each $i \in \{1, \dots, n+1\}$, let

$$\nu_i^{x,y}(\mathcal{D}) \in \operatorname*{arg\,min}_{\substack{1 \leq j \leq n+1 \\ j \neq i}} |\widetilde{X}_i - \widetilde{X}_j|$$

be chosen using a fixed measurable rule, common to all runs, for which the resulting score vector is permutation equivariant. For example, among the distance minimizers one may select the lexicographically smallest value $(\widetilde{X}_j, \widetilde{Y}_j)$, and then use the smallest index if that same pair occurs more than once. The final index tie-break may itself depend on the order of the indices. All indices remaining at that stage have the same pair value, so the order does not matter. Hence the induced score vector is permutation equivariant. Define the refitted leave-one-out 1-nearest-neighbor scores

$$R_i^{x,y}(\mathcal{D}) := \left|\widetilde{Y}_i - \widetilde{Y}_{\nu_i^{x,y}(\mathcal{D})}\right|.$$

Set

$$c_{\mathcal{D}}(x, y) := \sum_{i=1}^{n} \mathbf{1}\left\{R_i^{x,y}(\mathcal{D}) \geq R_{n+1}^{x,y}(\mathcal{D})\right\}, \qquad \widehat{p}_{\mathcal{D}}(x, y) := \frac{1 + c_{\mathcal{D}}(x, y)}{n+1},$$

and, for $m \in \{0, 1, \dots, n+1\}$, define the rank-cutoff map

$$\widehat{\mathcal{C}}_{\mathcal{D}}^{(m)}(x) := \{y \in \mathbb{R} : c_{\mathcal{D}}(x, y) \geq m\}.$$

Thus $\widehat{\mathcal{C}}_{\mathcal{D}}^{(0)} \equiv \mathbb{R}$ and $\widehat{\mathcal{C}}_{\mathcal{D}}^{(n+1)} \equiv \varnothing$.

Full statement of Proposition 8. Let $n \geq 2$, let P be a Borel probability law on $\mathbb{R}^2$ such that $P \ll \mathrm{Leb}^2$, and let $\mathcal{D}, \mathcal{D}'$ be independent $P^{\otimes n}$ samples. All events involving extensional equality of set-valued maps are understood in the completion of the relevant joint law. Equivalently, each zero-probability assertion below may be read as an outer-probability assertion. Then, for every $m \in \{1, \dots, n\}$,

$$\mathbb{P}\left[\widehat{\mathcal{C}}_{\mathcal{D}}^{(m)} \equiv \widehat{\mathcal{C}}_{\mathcal{D}'}^{(m)}\right] = 0,$$

where $\equiv$ means pointwise equality as set-valued maps on all of $\mathbb{R}$. More generally, let U be a shared random seed independent of $(\mathcal{D}, \mathcal{D}')$, and let $M = M(U) \in \{0, 1, \dots, n+1\}$ be measurable. If both runs use the same cutoff $M(U)$, then

$$\mathbb{P}\left[\widehat{\mathcal{C}}_{\mathcal{D}}^{(M)} \equiv \widehat{\mathcal{C}}_{\mathcal{D}'}^{(M)}\right] = \mathbb{P}[M \in \{0, n+1\}].$$

In particular, if $\mathbb{P}[M \in \{1, \dots, n\}] = 1$, then the exact-agreement probability is zero. For the usual convention

$$\widehat{\mathcal{C}}_{\mathcal{D},\alpha}(x) := \{y \in \mathbb{R} : \widehat{p}_{\mathcal{D}}(x, y) > \alpha\}, \qquad 0 < \alpha < 1,$$

one has

$$\widehat{\mathcal{C}}_{\mathcal{D},\alpha} = \widehat{\mathcal{C}}_{\mathcal{D}}^{(m_\alpha)}, \qquad m_\alpha := \lfloor \alpha(n+1) \rfloor.$$

Consequently,

$$\mathbb{P}\Big[\widehat{\mathcal{C}}_{\mathcal{D},\alpha} \equiv \widehat{\mathcal{C}}_{\mathcal{D}',\alpha}\Big] = 0 \quad \text{for every} \quad \alpha \in \left[\frac{1}{n+1}, 1\right),$$

whereas $0 < \alpha < 1/(n+1)$ gives the trivial always-$\mathbb{R}$ map. No condition on whether $\alpha(n+1)$ is an integer is needed.

E.1 Proof of Proposition 8

Proof. Because $P \ll \mathrm{Leb}^2$, both of its one-dimensional marginals are absolutely continuous and $P^{\otimes n} \ll \mathrm{Leb}^{2n}$. Hence, with probability one, the X_i's are pairwise distinct, every X_i has a unique nearest neighbor among the other training inputs, and the Y_i's are pairwise distinct. This holds because the exceptional events are contained in finite unions of affine hyperplanes of the forms

$$X_i = X_j, \qquad 2X_i = X_j + X_k, \qquad Y_i = Y_j.$$

Here the first and third forms use distinct pairs of indices, while the midpoint form uses distinct i, j, k. Let $\mathcal{G}$ denote this probability-one event. The stipulated tie rule defines the prediction map everywhere. On $\mathcal{G}$, it is irrelevant for the within-sample neighbors used below.

Fix $\mathcal{D} \in \mathcal{G}$, and define

$$J = J(\mathcal{D}) := \operatorname*{arg\,max}_{1 \le i \le n} X_i, \qquad N_i = N_i(\mathcal{D}) := \operatorname*{arg\,min}_{\substack{1 \le j \le n \\ j \ne i}} |X_i - X_j|,$$

$$d_i := |X_i - X_{N_i}|, \qquad T(\mathcal{D}) := \max_{1 \le i \le n} (X_i + d_i).$$

If $x > T(\mathcal{D})$, then $x > X_i$ for every i, because $d_i > 0$. The candidate's unique nearest neighbor is therefore J. In addition, for every training index i, $|x - X_i| = x - X_i > d_i$, so the candidate is not a nearest neighbor of any training observation. Consequently, for every $y \in \mathbb{R}$, augmentation by (x, y) leaves all training nearest neighbors unchanged and

$$R_i^{x,y}(\mathcal{D}) = r_i(\mathcal{D}) := |Y_i - Y_{N_i}|, \quad 1 \le i \le n, \qquad R_{n+1}^{x,y}(\mathcal{D}) = |y - Y_J|.$$

Write the training residuals in decreasing order, with multiplicities, as $r_{[1]}(\mathcal{D}) \ge \cdots \ge r_{[n]}(\mathcal{D})$, and fix $m \in \{1, \dots, n\}$. For every $a \ge 0$,

$$\#\{i : r_i(\mathcal{D}) \ge a\} \ge m \quad \iff \quad a \le r_{[m]}(\mathcal{D}).$$

It follows that, for every $x > T(\mathcal{D})$,

$$\widehat{\mathcal{C}}_{\mathcal{D}}^{(m)}(x) = \big[Y_J - r_{[m]}(\mathcal{D}),\, Y_J + r_{[m]}(\mathcal{D})\big].$$

Define the eventual far-right upper endpoint $B_m(\mathcal{D}) := Y_J + r_{[m]}(\mathcal{D})$. For completeness, set $B_m(\mathcal{D}) := 0$ on $\mathcal{G}^c$. This defines a Borel random variable: $\mathcal{G}$ is Borel, the unique finite argmax and argmin maps are Borel on $\mathcal{G}$, and absolute values and finite order statistics are Borel measurable.

We claim that $B_m(\mathcal{D})$ is atomless. Fix $t \in \mathbb{R}$. On $\mathcal{G}$, the order statistic $r_{[m]}$ equals r_i for at least one i, and $Y_i \ne Y_{N_i}$. Thus, for some $i \in \{1, \dots, n\}$ and $\sigma \in \{-1, 1\}$, $r_{[m]} = \sigma(Y_i - Y_{N_i})$. Therefore

$$\{B_m = t\} \cap \mathcal{G} \subseteq \bigcup_{\substack{1 \le j,i,k \le n,\ i \ne k \\ \sigma \in \{-1,1\}}} \{Y_j + \sigma(Y_i - Y_k) = t\}.$$

Each set in this finite union is a proper affine hyperplane. The coefficient vector of $Y_j + \sigma(Y_i - Y_k)$ is $e_j + \sigma(e_i - e_k)$, whose coordinate sum is 1, so it cannot be the zero vector. Absolute continuity of $P^{\otimes n}$ makes every such hyperplane null. Since $\mathbb{P}[\mathcal{G}^c] = 0$,

$$\mathbb{P}[B_m(\mathcal{D}) = t] \le \mathbb{P}[\mathcal{G}^c] + \sum_{\substack{1 \le j,i,k \le n,\ i \ne k \\ \sigma \in \{-1,1\}}} \mathbb{P}[Y_j + \sigma(Y_i - Y_k) = t] = 0$$

for every $t \in \mathbb{R}$, proving atomlessness. This argument does not require the residuals to be distinct. In particular, if $N_i = k$ and $N_k = i$, then the structural duplicate $r_i = r_k = |Y_i - Y_k|$ is handled without modification.

Now let $\mathcal{D}, \mathcal{D}'$ be independent. Then $B_m(\mathcal{D})$ and $B_m(\mathcal{D}')$ are independent and atomless, whence $\mathbb{P}[B_m(\mathcal{D}) = B_m(\mathcal{D}')] = 0$. On the probability-one event $\{\mathcal{D} \in \mathcal{G},\ \mathcal{D}' \in \mathcal{G}\}$, take the common input $x_* := 1 + \max\{T(\mathcal{D}), T(\mathcal{D}')\}$. If $\widehat{\mathcal{C}}^{(m)}_{\mathcal{D}} \equiv \widehat{\mathcal{C}}^{(m)}_{\mathcal{D}'}$, then their interval values at x_* are equal, so their upper endpoints are equal. Writing $E_m := \{\widehat{\mathcal{C}}^{(m)}_{\mathcal{D}} \equiv \widehat{\mathcal{C}}^{(m)}_{\mathcal{D}'}\}$, we therefore have

$$E_m \subseteq \{B_m(\mathcal{D}) = B_m(\mathcal{D}')\} \cup \{\mathcal{D} \notin \mathcal{G}\} \cup \{\mathcal{D}' \notin \mathcal{G}\}.$$

The right-hand side is a measurable null set. Hence E_m has outer probability zero and belongs to the completed product sigma-field, where it has probability zero. This proves the fixed-m assertion without requiring a separate function-space sigma-field.

For the shared-seed claim, write

$$E_M := \left\{\widehat{\mathcal{C}}^{(M)}_{\mathcal{D}} \equiv \widehat{\mathcal{C}}^{(M)}_{\mathcal{D}'}\right\}, \qquad A := \{M \in \{0, n+1\}\}.$$

The two endpoint maps agree identically, so $A \subseteq E_M$. In addition,

$$E_M \setminus A \subseteq \bigcup_{m=1}^{n} \left(\{M = m\} \cap E_m\right) \subseteq \bigcup_{m=1}^{n} E_m.$$

After lifting the E_m's to the joint space carrying $(U, \mathcal{D}, \mathcal{D}')$, the last union is a finite null set. Thus E_M differs from A only by a subset of a measurable null set, and hence, in the completed joint law,

$$\mathbb{P}[E_M] = \mathbb{P}[A] = \mathbb{P}[M \in \{0, n+1\}].$$

This argument avoids any need to factor probabilities of completion-measurable map-identity events.

Finally, since $c_{\mathcal{D}}(x, y)$ is integer-valued,

$$\widehat{p}_{\mathcal{D}}(x, y) > \alpha \quad \Longleftrightarrow \quad c_{\mathcal{D}}(x, y) > \alpha(n+1) - 1 \quad \Longleftrightarrow \quad c_{\mathcal{D}}(x, y) \geq \lfloor \alpha(n+1) \rfloor.$$

This equivalence holds whether or not $\alpha(n+1)$ is an integer. For $\alpha \in [1/(n+1), 1)$, the cutoff belongs to $\{1, \ldots, n\}$, so the zero-agreement result applies. For $0 < \alpha < 1/(n+1)$, the cutoff is zero and every candidate is accepted, yielding the always-$\mathbb{R}$ map. This completes the proof. □

F Experimental details

In this appendix, ReCal denotes seeded rounding in score units (Algorithm 1 with a shared offset and G the identity) at the width stated with each result. It is the rule of the in-pool, language-model and answer-cache experiments. The *plug-in width* is $\sqrt{2\alpha(1-\alpha)/n}/(\hat{f}\rho)$, with the pilot density estimate $\hat{f}$ described under "Scores, jitter and pilot density" below, and "$\hat{\kappa}$=1.5" multiplies it by 1.5 to allow for a density ratio. The *conservative width* is $2\hat{\kappa}B_n^+/(\hat{f}_{\mathrm{L}}\rho)$ with $\hat{\kappa}$=1.5 and the split-pilot density estimate $\hat{f}_{\mathrm{L}}$ of Proposition 6, scaled down by $1/(1+c)$ with c=.1, the score-unit width whose mismatch guarantee holds under the margin condition. The *pilot-centred grid* is the seedless score-unit grid of Proposition 7, of plug-in width with a cell centred at the pilot's $(1-\alpha)$-quantile. It differs from the centred grid of the main text, which lives in coverage units. The pilot-only threshold uses the *conservative rank*, the smallest pilot rank whose coverage is at least $1-\alpha$ with probability .95 (the training-conditional rank at δ=.05). The other rules carry the names of the main text, and the premium of a rule is relative to standard split conformal on the same pairs.

F.1 Protocols, estimators and uncertainty

Master protocol. Every table in the main text and every table below marked "master protocol" uses one design, scaled to the pool. The pool's fixed evaluation subsample (seed 9001) is the held-out test split on which coverage and sizes are measured. A public pilot is drawn from the remaining points, and the rest is split into two disjoint analyst halves. Each analyst in a pair draws n points from its own half without replacement. Pilot redraws re-split the pool. The sizes are as follows. Image pools of 50,000: test 10,000, pilot 2,000, halves 19,000, n=15,000 (79% of a half), 1,000 pairs over 20 redraws. CIFAR-100 released probabilities (N=10,000): test 2,500, pilot 937, halves 3,281, n=2,460. Fine-tuned CIFAR-100 logits (N=20,000): test 5,000, pilot 1,875, halves 6,562 (one point unused), n=4,921 (75% of a half). Language models with N=10,000 contexts (GPT-2, Pythia-1.4B, Qwen2.5-1.5B): test 2,000, pilot 1,000, halves 3,500, n=2,625. Llama-3.2-1B (N=20,000): test 2,000, pilot 2,000, halves 8,000, n=6,000. The 40,000-context pools: test 2,000, pilot 2,000, halves 18,000, n=13,500. Language models use 1,000 pairs over 10 redraws.

Estimators. For a seeded rule on a common score grid of width β, identity is integrated exactly over the shared offset, $1 - \min\{|\widetilde{\tau}_A - \widetilde{\tau}_B|/\beta, 1\}$ per pair (Lemma 2). For seeded rounding in coverage units it is integrated in level units, and rounded levels at or above 1 all map to $+\infty$: a pair with levels $g_A \le g_B$ mismatches with probability $\min\{(\min\{g_B, 1\} - g_A)_+/b, 1\}$. The pilot map G is the pilot's empirical distribution function interpolated linearly through the points $(Z_{(i)}, (i - \frac{1}{2})/n_p)$, where $Z_{(1)} \le \cdots \le Z_{(n_p)}$ are the pilot scores, and ramped to 0 and 1 over intervals of a fixed small length outside the pilot's range, so that $\|G - \widehat{F}_{n_p}\|_\infty \le 1/(2n_p)$, that is $\lambda = \frac{1}{2}$. It is continuous and strictly increasing where $0 < G < 1$, so for rounded levels below 1 equal rounded levels are equivalent to equal thresholds, and levels at or above 1 deploy $+\infty$. (In `v51_master.py` the centred grid's inverse maps a level at or above 1 to a point just above the largest pilot score, where the paper's convention is $+\infty$, and the seeded rules' coverage and size go through `exact_offsets.py`, which treats such levels as $+\infty$. No rounded level reached 1 in its runs, so neither convention binds.) For seedless rules, identity is the observed equality of the two deployed thresholds. Coverage is the held-out fraction of test scores at or below the deployed threshold. The held-out split has its own sampling error of about $\pm.003$ at a fixed level ($\pm.006$ on the 2,500-point CIFAR-100 split) and of about $\pm.001$ on the difference between two levels .013 apart, so it does not cancel in the measured inflations. Size is the mean over test inputs of $\max\{|\mathcal{C}(x)|, 1\}$, the convention that a system deploys the top-ranked label when the raw set is empty. Coverage is that of the raw set, the object of the guarantee, and does not count this label (counting it would add about .055 to the standard APS coverage on ImageNet). At the standard APS cutoff the raw set is empty for 5.5% of ImageNet test inputs, and the convention adds .06 to the mean size of 34.6. For RAPS the fraction is 2.7% and the addition .027 on a mean of 5.5. Every rule of a table shares the convention, and a rule with a larger threshold has fewer empty raw sets, so the convention changes $1 +$ premium by a factor within .2% (APS) and .5% (RAPS) of one: at most .2 points at $\rho = .1$, about 3 points at the saturated APS premiums near +2,000% and about 60 points at the RAPS coverage-unit premium of +12,752%. Sizes use 52 pairs from the first four redraws (`v51_master.py`, standard size 34.59 on ImageNet and 14.80 on the synthetic pool), 200 pairs from all redraws (`v6_regimes.py`, standard size 34.55 and 14.87) or all 1,000 pairs (`v8_extras.py`, Tables 2 and S2, standard size 34.58 and 14.87), so the last digit of a premium can differ between the scripts. For seeded rules on the image pools, coverage, size and the undercoverage fraction are integrated exactly over the shared offset (`exact_offsets.py`). Whatever the raw value x, the deployed value is $x + W$ with W uniform on $[0, b)$, and coverage and size are step functions of the threshold, so the integral is a finite sum over the sorted test scores. The language-model and fine-tuned CIFAR-100 runs, whose logits are not in the archive, average size over four fixed offsets and coverage over 16, and the answer-cache run averages both over four offsets. The four-offset size estimate is accurate when the phase of the raw cutoff relative to the grid varies between pairs, and it is biased when all pairs share almost the same phase and part of a cell deploys the full label set. On ImageNet at ρ=.01, on the 200 size pairs of `v6_regimes.py`, it gives +1,535% where the exact value is +1,994% for seeded coverage-unit rounding, and +1,876% where it is +1,831% for the score-unit rule. At ρ=.1 the four-offset estimate of the score-unit premium is within 2% of the exact value for APS and within 6% for RAPS on the image pools. Unless a caption says otherwise, $\pm$ is a standard error clustered by pilot redraw, because pairs within one redraw share the pilot and the split into halves. Sampled-offset streams report binomial standard errors, and zero failures in m independent pairs give the one-sided 95% bound $1 - 0.05^{1/m} \approx 3/m$.

Finite pools and the i.i.d. theory. The pool experiments sample without replacement from a finite pool, whose empirical law has atoms of size $1/N$. Let N now count the points outside the test split and the pilot, split into complementary halves from each of which one analyst draws n. To first order in the Bahadur representation, the difference of the two analysts' cutoffs then has the variance of i.i.d. sampling from the pool (up to a factor $N/(N-1)$), so the seeded measurements, which depend on a pair only through this difference (Lemma 2), are faithful to the i.i.d. theory. Each analyst's own cutoff, however, has $1 - n/N$ times the i.i.d. variance around the pool quantile (.61 in the master protocol), and the two cutoffs are negatively correlated, with correlation $-n/(N-n) = -.65$ to first order ($-.60$ on the CIFAR-100 pool). Conditionally on one split, most of the difference is the fixed offset between the two halves, which is why standard errors are clustered by redraw. Seedless identity depends on where each cutoff falls, so we checked the seedless certificate in a second design on the same 100 pilot redraws $\times$ 20 pairs as Table S5: both analysts resample i.i.d., with replacement, from the pool left after the test split and the pilot. At n_p=250/500/1,000/2,000/4,000 the i.i.d. identities

are .968/.969/.954/.980/.938 on ImageNet and .990/.959/.960/.966/.911 on the synthetic pool, against certificates .914/.904/.904/.903/.900. Complementary halves give identities at least as high in 9 of the 10 settings, by up to .037. The largest gap in standard-error units is 2.2, at synthetic n_p=2,000. Every point estimate in both designs exceeds its certificate. The smallest margin, .010 at synthetic n_p=4,000, is within one standard error. The fresh-draw check below instantiates the i.i.d. assumptions exactly. The earlier in-pool protocol, used only in tables so marked, resamples with replacement from the whole pool and measures coverage on the pool. Repeated indices there give standard calibration small nonzero identities ($\leq$ .02 on the image pools, up to .05 on digits and .04 on the in-pool language-model runs).

Scores, jitter and pilot density. The APS score is $s_i = V_{i,o_i} - u_i p_{i,(o_i)}$, with V the cumulative sorted probabilities, o_i the rank of the true label, and $u_i \in [0, 1)$ a fixed per-instance pseudo-random number shared by all analysts (Remark 1). RAPS adds $\lambda_{\text{reg}} \max(0, o - k_{\text{reg}})$ with $(k_{\text{reg}}, \lambda_{\text{reg}}) = (5, 0.05)$. A finite-range pseudo-random jitter has atoms, so the pool experiments do not instantiate the continuous theory exactly. The pilot density $\hat{f}$ is the pilot mass in the band between its $(1 - \alpha \mp .05)$-quantiles divided by the band's width, a scale-free window holding about $0.1n_p$ pilot points (200 of 2,000). The in-pool experiments below used a fixed window of half-width .02 (.05 for the clinical sites), which suits the unit APS scale but not the RAPS scale of a 50k–152k vocabulary. The language-model runs therefore use the scale-free window.

Pools. Seeds are fixed: pool 20260808, instance jitter 777, analyst resampling 424242, test split 9001. *Synthetic image pool*: N=50,000, K=1,000, uniform labels, per-example margin $\mu_i = \text{softplus}(m + \xi_i)$ with $\xi_i \sim \mathcal{N}(0, 1)$ and m tuned by bisection to a top-1 of 0.78, logits $z_i = \mu_i e_{y_i} + \mathcal{N}(0, I_K)$. The achieved top-1 is 79.0% and the top-5 is 89.9% (independent-coordinate logits reduce the top-5 mass relative to real networks). *Real ImageNet*: the released validation logits of a ResNet-50 (Northcutt et al., 2021), pre-softmax (range -11 to 28), top-1 .727, top-5 .903. *CIFAR-100, released probabilities*: the cross-validated predicted probabilities of the CIFAR-100 test set released with the same project (Northcutt et al., 2021) (N=10,000, K=100, rows summing to one, top-1 .693, top-5 .898), passed to the score builder as log-probabilities, which its softmax maps back to the released probabilities, with held-out test 2,500, pilot 937, halves of 3,281 and n=2,460 by the split rule of `v51_master.py` for pools below 40,000 points (a quarter of the pool as test split, a pilot of at most 2,000 points and at most an eighth of the rest, two halves of the remainder with one point unused when odd, and n equal to three quarters of a half). *Digits*: scikit-learn digits (Pedregosa et al., 2011), multinomial logistic regression (C=0.5) fitted on a stratified half, pool = the other 899 images (accuracy .957).

F.2 Coverage-unit and seedless experiments

Coverage-unit rounding at the implemented widths. Table S1 and the ImageNet rows quoted in its caption run rounding in coverage units at three implemented widths beside the score-unit rules, on the same pilots, pairs and test split as Table 2. At $(n, n_p, \rho) = (15{,}000, 2{,}000, .1)$, Theorem 2(iii-b) with $\lambda = \frac{1}{2}$ certifies mismatch at most $(D_n + 1/n_p)/b$. This is .094 at $b_0 = \sqrt{2}\sigma_F/\rho$, .046 at $2B_n^+/\rho$, and .118 at the asymptotically sharp $(2/\sqrt{\pi})\sigma_F/\rho$, which is certified only for the oracle map ($D_n/b = .100$). Measured mismatch is .060/.089, .029/.043 and .075/.112 (synthetic/ImageNet). The centred grid at b_0 is certified at identity .976 by Theorem 3(ii) and measured at 1.000 on both pools. On ImageNet, rounding in coverage units at b_0 costs +21% on the 52 size pairs of `v51_master.py`, the same as the uncertified score-unit plug-in width on those pairs, and Table 2 gives +22% for both on all 1,000 pairs, with the coverage-unit rule at the certified width $(D_n + 1/n_p)/\rho = .0326$. The conservative score-unit width, the only score-unit row of certificate form, costs +134%.

Widths in the ρ-sweep. The seedless widths are the smallest b on a lattice of step $1/(4n_p)$ whose certificate is at least $1 - \rho$: b = .0265/.0306/.0396, with certificates .9025/.9522/.9911, at ρ = .1/.05/.01. The seeded coverage-unit widths are $(D_n + 1/n_p)/\rho$ = .0326/.0653/.3264, and the pilot-floor ranks are j = 1,822/1,826/1,835, with exact identities .9122/.9507/.9900 and mean coverages .9107/.9126/.9171 under i.i.d. sampling. At ρ=.01 the seeded width exceeds α by a factor of three, so many rounded levels reach 1 and deploy the full label set. This is why its premium is large on both pools. The score-unit plug-in width scales as $1/\rho$ in the same way. With $b \propto \sqrt{\log(1/\rho)}$,

Table S1: Synthetic image pool, master protocol at ρ=.1 (same design as Table 2, from `v51_master.py` with sizes on 52 pairs, and the seeded rules at their implemented widths, which for coverage units are $b_0 = \sqrt{2}\sigma_F/\rho$ and two alternatives). The pilot-floor and pilot-only rows come from 200 pairs over all 20 redraws, whose standard size is 14.87, and their premiums are relative to it. The same rules at the certified widths of Section 7, with sizes on all 1,000 pairs, are in Table S2. On real ImageNet, the rows omitted from Table 2 are seeded score-unit rounding at the conservative width (coverage .9528, size 80.88, identity .978 ± .003, +134%) and seeded coverage-unit rounding at $b = 2B_n^+/\rho$ (coverage .9366, size 56.23, identity .957 ± .006 against the certificate .954, +63%).

Rule	Shared	Coverage	Size	Identity	Certified	Premium
Standard APS	none	.8988	14.80	.000		+0%
Seeded, score units, plug-in width	seed, $\hat{f}$	.9153	18.86	.934 ± .008		+27%
Seeded, score units, conservative width	seed, $\hat{f}_{\mathrm{L}}$	.9339	25.33	.974 ± .003		+71%
Seeded, coverage units, $b = b_0$	seed, pilot	.9176	19.81	.940 ± .008	$\geq$ .906	+34%
Seeded, coverage units, $b = 2B_n^+/\rho$	seed, pilot	.9367	30.50	.971 ± .004	$\geq$ .954	+106%
Seeded, coverage units, $b = (2/\sqrt{\pi})\sigma_F/\rho$	seed, pilot	.9138	18.62	.925 ± .010	$\geq$ .882	+26%
Centred grid, $b = b_0$	pilot	.9146	17.87	1.000 ± .000	$\geq$ .976	+21%
Pilot floor, rank 1,822	pilot	.9077	16.85	.846 ± .076	= .912	+13%
Pilot-only threshold, conservative rank	pilot	.9077	16.89	1.000	= 1	+14%
Standard RAPS	none	.9017	6.92	.000		+0%
Seeded RAPS, score units, plug-in width	seed, $\hat{f}$	.9216	8.83	.957 ± .007		+28%

the centred grid's coverage rises only from .914 to .919 (ImageNet) as ρ falls from .1 to .01, and the pilot floor's from .911 to .917.

RAPS and a second real pool. Table S3 repeats the ImageNet sweep with the RAPS score on the same pilots and pairs, and Table S4 runs the same protocol on a second real pool, the released cross-validated predicted probabilities of the CIFAR-100 test set (Northcutt et al., 2021), at $(n, n_p) = (2{,}460, 937)$, with the certificates recomputed for these sizes. In all four pool–score combinations the seeded coverage-unit identity is at or above $1 - \rho$ at every ρ, and the centred grid is above its certificate or within .5 clustered standard errors of it (ImageNet RAPS at $\rho = .05$ and .02, .949 ± .049 against .952 and .961 ± .039 against .980, and CIFAR-100 APS at $\rho = .1$ and .05, within .15). The pilot floor is within 1.5 standard errors of its exact value with RAPS on ImageNet, at most .8 below and up to 2.3 above it with APS on CIFAR-100 (.995 ± .005 against .984 at $\rho = .02$), and with RAPS on CIFAR-100 it lies 1.1 to 2.1 standard errors below the exact value at $\rho \geq .1$. The complementary halves make the two cutoffs negatively dependent (correlation $-n/(N-n)$, −.60 on CIFAR-100 and −.65 on ImageNet) and reduce their spread, and a Gaussian model with both effects raises the floor's exact identity by at most .006 on ImageNet and .015 on CIFAR-100 (`finite_pool_model.py`), so the finite pool does not produce identities below the exact values, and the deviations are within the variability of 20 pilot redraws. Under fresh i.i.d. draws at $(2{,}460, 937)$ the floor is within one standard error of its exact value (Table S8). The smaller CIFAR-100 sample makes every width larger: the seeded coverage-unit width $(D_n + 1/n_p)/\rho$ is .079 at $\rho = .1$ and exceeds α from $\rho = .05$, and the plug-in score-unit width already sends most APS thresholds to the full label set at $\rho = .1$ (+1,140%), whereas the floor and the grid cost +20% and +29% at $\rho = .1$ and +37% and +45% at $\rho = .01$ with APS, against +18% for the pilot-only threshold. With RAPS the standard sets have 3.8 labels, and the seedless premiums are +49% and +78% at $\rho = .1$ and +82% and +102% at $\rho = .01$, against +101% and +797% for seeded score-unit rounding and +44% for the pilot-only threshold.

Pilot floor. The rule of Proposition 1 deploys $\max\{\widetilde{\tau}, Z_{(j)}\}$ with $Z_{(j)}$ the j-th smallest jittered score of the same 2,000-point pilot that the other rules use, at the smallest rank j whose exact identity $\mathbb{E}[H_n(V_j)^2]$ is at least $1 - \rho$ (`certificates.py`). Its identity is the observed equality of the two deployed thresholds, which is 1 when both raw cutoffs lie at or below $Z_{(j)}$ and 0 otherwise, so it is decided almost entirely by the pilot: for a given pilot the two analysts agree or disagree together, and the between-pilot standard deviation of the per-pilot identity is .21 on ImageNet and .34 on the synthetic pool at ρ=.1. With 20 pilots per setting the clustered standard errors are therefore .035 to .076 at ρ=.1 and .05, and there the measured identities of Table S2 lie within one standard error of the exact values on both pools, above them on ImageNet (.949 and .964 against .912 and .951) and below them on the synthetic pool (.846 and .907). At ρ=.01 the anchor at rank 1,835 lies above both raw cutoffs for all 20 pilots and all 1,000 pairs on both pools, so the measured identity is

Table S2: **Identity, coverage and premium as ρ decreases** (master protocol, both pools, 1,000 pairs per setting with sizes on all pairs, the values plotted in Figure 2). Each rule uses the width or rank chosen for that ρ: the plug-in width for seeded score-unit rounding, which has no certificate, $(D_n + 1/n_p)/\rho$ for seeded coverage-unit rounding, the smallest certified width b for the centred grid and the certified rank for the pilot floor. "Certified" gives the certified threshold identity, the same for both pools (a lower bound for the seeded and centred rules, the exact value for the pilot floor), and $\pm$ is clustered by pilot redraw. Premiums are relative to standard split conformal on the same pairs. The last rows do not depend on ρ. Replicable binary search with an upward-shifted target, floored at the local cutoff, which makes it exactly valid, has identity .39/.64/.80 at +11/38/107% (coverage .9112/.9263/.9496) on ImageNet and .48/.69/.85 at +15/48/235% (.9087/.9244/.9541) on the synthetic pool for $\varepsilon = .02/.05/.1$ (with sizes on 200 pairs, `v6_regimes.py` gives +11/37/106% on ImageNet), the floored search with a symmetric target has identity .32 at +5% on ImageNet for $\varepsilon = .05$, and the unfloored symmetric search has identity .40/.64/.75 at +1/2/3% (coverage .9024/.9022/.8993, below .90 on 32/39/46% of pairs) on ImageNet and .49/.69/.78 at +0/2/11% on the synthetic pool. Table S9 reports the same searches from `v51_master.py`, whose pairs and search randomness differ, within .04. The deterministic grid without a seed, at the plug-in width of the score-unit rule, has identity .698/.820/.841/.955/.989/.966/1.000 on ImageNet at premiums +4/7/11/28/55/1,510/2,372% along the same ρ grid, a value fixed by the position of the pool's quantile within a cell (Figure S3).

			Real ImageNet			Synthetic pool		
Rule	ρ	Certified	Identity	Coverage	Premium	Identity	Coverage	Premium
Seeded, score units	.5		.633 ± .054	.9055	+4%	.672 ± .039	.9026	+5%
	.3		.775 ± .035	.9076	+7%	.802 ± .024	.9049	+9%
	.2		.850 ± .023	.9101	+10%	.868 ± .016	.9076	+13%
	.1		.925 ± .012	.9176	+22%	.934 ± .008	.9153	+28%
	.05		.962 ± .006	.9328	+55%	.967 ± .004	.9282	+60%
	.02		.985 ± .002	.9674	+868%	.987 ± .002	.9522	+186%
	.01		.992 ± .001	.9837	+1,830%	.993 ± .001	.9711	+515%
Seeded, cov. units	.5	≥.500	.556 ± .057	.9051	+3%	.686 ± .044	.9028	+5%
	.3	≥.700	.717 ± .041	.9071	+6%	.811 ± .027	.9050	+9%
	.2	≥.800	.811 ± .028	.9096	+9%	.874 ± .018	.9078	+13%
	.1	≥.900	.906 ± .014	.9174	+22%	.937 ± .009	.9166	+30%
	.05	≥.950	.953 ± .007	.9335	+58%	.968 ± .004	.9336	+85%
	.02	≥.980	.981 ± .003	.9699	+1,193%	.987 ± .002	.9703	+2,865%
	.01	≥.990	.991 ± .001	.9850	+1,992%	.994 ± .001	.9851	+4,746%
Centred grid	.5	≥.513	.663 ± .081	.9071	+6%	.786 ± .066	.9049	+9%
	.3	≥.707	.838 ± .067	.9083	+8%	.872 ± .057	.9081	+14%
	.2	≥.807	.897 ± .065	.9104	+11%	.929 ± .033	.9104	+18%
	.1	≥.903	.977 ± .017	.9137	+15%	.961 ± .031	.9126	+23%
	.05	≥.952	.999 ± .001	.9155	+18%	.991 ± .007	.9143	+26%
	.02	≥.980	1.000 ± .000	.9178	+21%	1.000 ± .000	.9150	+26%
	.01	≥.991	1.000 ± .000	.9194	+24%	1.000 ± .000	.9174	+30%
Pilot floor	.5	=.516	.414 ± .101	.9040	+2%	.353 ± .107	.9009	+3%
	.3	=.708	.683 ± .098	.9062	+5%	.562 ± .108	.9032	+6%
	.2	=.819	.878 ± .067	.9084	+8%	.713 ± .098	.9048	+9%
	.1	=.912	.949 ± .048	.9113	+12%	.846 ± .076	.9077	+13%
	.05	=.951	.964 ± .035	.9132	+14%	.907 ± .058	.9096	+17%
	.02	=.982	1.000 ± .000	.9158	+18%	.949 ± .050	.9130	+22%
	.01	=.990	1.000 ± .000	.9173	+21%	1.000 ± .000	.9145	+25%
Standard split conformal			.000	.9018	+0%	.000	.8988	+0%
Pilot only, conservative rank			1.000	.9117	+12%	1.000	.9077	+14%
Pilot only, plain rank			1.000	.9015	0%	1.000	.8951	−4%

1.000 against the exact .990. Table S5 repeats the comparison with 100 pilots per setting, and the fresh-draw check below with 5,000 independent pilots per law. The floor's mean coverage under i.i.d. sampling, $\mathbb{E}\max\{U, V_j\}$, is .9107 at ρ=.1, and the measured held-out coverage is .9113 on ImageNet and .9077 on the synthetic pool, whose standard cutoff covers .8988 on the held-out split.

Pilot size. Every measured identity of the centred grid and of the seeded rule in Table S5 exceeds its certificate, and the pilot floor's measured identities lie within 1.2 standard errors of their exact values in all ten settings, below them in four of the five ImageNet settings (by at most .023, at n_p=500) and above them in every synthetic setting (by at most .025). As n_p grows the certified width of the grid shrinks and the size with it, from 48.3 at n_p=500 to 38.6 at n_p=4,000 on ImageNet (standard: 34.6), and the floor's size falls from 42.9 to 37.7. The seeded width $(D_n + 1/n_p)/\rho$ tends to $D_n/\rho = .028$. Small pilots increase the set size of the grid in the tail of the cutoff distribution. At n_p=250 the

Table S3: **RAPS on real ImageNet as ρ decreases** (master protocol, the same pilots and pairs as Table S2, RAPS score, standard size 5.54, sizes on all $1{,}000$ pairs). Columns as in Table S2. Replicable binary search floored at the local cutoff has identity $.66/.78/.87$ at $+6/20/666\%$ (coverage $.9106/.9268/.9641$) for $\varepsilon = .02/.05/.1$, the unfloored symmetric search $.67/.79/.86$ at $+1/2/4\%$ (coverage $.9011/.9010/.9017$), and the deterministic grid without a seed $.692/.810/.864/.944/.950/.961/.961$ along the ρ grid at premiums $+2/3/7/18/35/83/144\%$.

Rule	ρ	Certified	Identity	Coverage	Premium
Seeded, score units	.5		$.671 \pm .049$	.9029	$+2\%$
	.3		$.802 \pm .030$	.9058	$+4\%$
	.2		$.868 \pm .020$	.9093	$+6\%$
	.1		$.934 \pm .010$	.9193	$+13\%$
	.05		$.967 \pm .005$	.9316	$+33\%$
	.02		$.987 \pm .002$	.9440	$+85\%$
	.01		$.993 \pm .001$	.9532	$+166\%$
Seeded, cov. units	.5	$\geq.500$	$.577 \pm .056$	.9019	$+2\%$
	.3	$\geq.700$	$.732 \pm .040$	.9041	$+3\%$
	.2	$\geq.800$	$.821 \pm .027$	.9069	$+4\%$
	.1	$\geq.900$	$.911 \pm .014$	.9152	$+9\%$
	.05	$\geq.950$	$.955 \pm .007$	.9317	$+61\%$
	.02	$\geq.980$	$.982 \pm .003$	.9692	$+7{,}553\%$
	.01	$\geq.990$	$.991 \pm .001$	.9846	$+12{,}752\%$
Centred grid	.5	$\geq.513$	$.748 \pm .081$	.9045	$+3\%$
	.3	$\geq.707$	$.872 \pm .062$	.9061	$+4\%$
	.2	$\geq.807$	$.872 \pm .060$	.9069	$+4\%$
	.1	$\geq.903$	$.918 \pm .055$	.9101	$+6\%$
	.05	$\geq.952$	$.949 \pm .049$	.9137	$+8\%$
	.02	$\geq.980$	$.961 \pm .039$	.9159	$+10\%$
	.01	$\geq.991$	$1.000 \pm .000$	.9161	$+10\%$
Pilot floor	.5	$=.516$	$.368 \pm .103$	.9009	$+1\%$
	.3	$=.708$	$.577 \pm .109$	.9028	$+2\%$
	.2	$=.819$	$.712 \pm .095$	.9046	$+3\%$
	.1	$=.912$	$.855 \pm .073$	.9073	$+5\%$
	.05	$=.951$	$.947 \pm .050$	.9096	$+6\%$
	.02	$=.982$	$.953 \pm .047$	.9124	$+7\%$
	.01	$=.990$	$.961 \pm .039$	.9137	$+8\%$
Standard split conformal			.000	.8982	$+0\%$
Pilot only, conservative rank			1.000	.9076	$+5\%$
Pilot only, plain rank			1.000	.8957	-1%

certified width is $b = .069$, so the grid point above the central cell is at a level of at least 1, and any analyst whose transformed cutoff lies above the central cell deploys the full label set. This is why the mean size of the grid at n_p=250 is 88.4 on ImageNet, against 48.4 for the floor, whose anchor at rank 232 of 250 has expected level $232/251 = .924$ (measured floor coverage $.926$) and needs no cell above it.

Shifted pilot: design. S is the set of 500 ImageNet classes whose within-class 0.9-quantile of the APS score is below the median over classes, a label-shifted population with smaller scores. The classes are chosen on the pool to produce a controlled shift. The public pilot (n_p=2,000) contains a fraction π_S of images from S and $1 - \pi_S$ from the other classes, and π_S=.5 is close to the analysts' own mix. For each of the 40 redraws, the pilot is taken from a reservoir of $5{,}000$ points of the pool outside the test split, and the two analysts' halves ($17{,}500$ points each, n=15,000) are a random split of the remaining $35{,}000$. The halves and pairs are the same for every π_S, so only the pilot changes from row to row, and the analysts and the test split keep the original population. The coverage-unit widths and the floor rank are fixed at the values certified for a representative pilot, because the analysts cannot detect the shift from the pilot, and the score-unit plug-in width is recomputed from each pilot's $\hat{f}$. For each pilot we also compute the lattice prediction of Theorem 3(iv), $\sum_m \pi_m(G)^2$, using the analysts' pool distribution function to map the pilot's cell boundaries to coverage levels. It fits the per-pilot measured identities with R^2=.67 and mean absolute error $.065$ for the positive-gap rows ($\pi_S \geq .5$) and R^2=.60 and mean absolute error $.111$ for the negative-gap rows ($\pi_S \leq .4$, run separately by `v8_extras.py` with the same design, seeds and pairs), shown together in Figure 3(b). Sizes in Table S6 use 200 of the $1{,}000$ pairs per row, and premiums are relative to the master protocol's standard size (34.58), which the raw local cutoff's size in this design (34.5) matches to within $.1$. For the pilot floor we record the anchor level $F(Z_{(j)})$ under the analysts' population law, where F is the distribution function of the $35{,}000$ analyst points, the pool outside the test split and the pilot reservoir.

Table S4: **A second real pool: CIFAR-100** (the released cross-validated predicted probabilities of the CIFAR-100 test set (Northcutt et al., 2021), N=10,000, K=100, top-1 .693, master protocol with held-out test 2,500, pilot n_p=937, disjoint halves of 3,281 and n=2,460 per analyst, 20 pilot redraws $\times$ 50 pairs, sizes on all pairs, `v8_extras.py --pyx`). Same columns and conventions as Table S2, with the certificates recomputed at $(n, n_p) = (2{,}460, 937)$: the certified widths of the centred grid are .023/.030/.036/.045/.051/.059/.065 and the floor ranks 848/853/856/860/864/868/870 of 937 for $\rho = .5/.3/.2/.1/.05/.02/.01$, and the seeded coverage-unit widths $(D_n + 1/n_p)/\rho$ reach .16 at $\rho = .05$, so with APS the score-unit rule deploys most of the label set from $\rho = .1$ (+1,140%) and the coverage-unit rule from $\rho = .02$ (+1,601%). With APS, replicable binary search floored at the local cutoff has identity .26/.48/.70 at +12/32/83% for $\varepsilon = .02/.05/.1$ (with RAPS .28/.48/.63 at +41/86/327%), the unfloored symmetric search .26/.42/.58 at $-0/+0/-1$% with coverage .896/.895/.891, and the deterministic grid without a seed .555/.835/.852/.988/.994/1.000/1.000 along the ρ grid at premiums +12/22/259/970/1,681/1,786/1,889%.

			APS			RAPS		
Rule	ρ	Certified	Identity	Coverage	Premium	Identity	Coverage	Premium
Seeded, score units	.5		.592 ± .048	.9078	+12%	.897 ± .035	.9092	+41%
	.3		.749 ± .032	.9178	+48%	.938 ± .021	.9125	+51%
	.2		.832 ± .021	.9345	+283%	.959 ± .014	.9162	+63%
	.1		.916 ± .011	.9667	+1,140%	.979 ± .007	.9260	+101%
	.05		.958 ± .005	.9833	+1,612%	.990 ± .004	.9401	+177%
	.02		.983 ± .002	.9933	+1,896%	.996 ± .001	.9640	+409%
	.01		.992 ± .001	.9967	+1,990%	.998 ± .001	.9795	+797%
Seeded, cov. units	.5	$\geq$.500	.566 ± .056	.9059	+10%	.623 ± .056	.9072	+30%
	.3	$\geq$.700	.722 ± .040	.9120	+18%	.747 ± .048	.9125	+48%
	.2	$\geq$.800	.814 ± .027	.9187	+29%	.830 ± .032	.9193	+74%
	.1	$\geq$.900	.907 ± .014	.9396	+84%	.915 ± .016	.9403	+217%
	.05	$\geq$.950	.954 ± .007	.9686	+876%	.958 ± .008	.9689	+1,189%
	.02	$\geq$.980	.981 ± .003	.9874	+1,601%	.983 ± .003	.9876	+1,987%
	.01	$\geq$.990	.991 ± .001	.9937	+1,843%	.992 ± .002	.9938	+2,252%
Centred grid	.5	$\geq$.524	.740 ± .070	.9098	+15%	.596 ± .069	.9111	+42%
	.3	$\geq$.707	.800 ± .067	.9139	+20%	.767 ± .061	.9157	+61%
	.2	$\geq$.803	.816 ± .066	.9163	+24%	.845 ± .057	.9189	+71%
	.1	$\geq$.907	.899 ± .055	.9202	+29%	.942 ± .031	.9200	+78%
	.05	$\geq$.951	.946 ± .037	.9233	+34%	.981 ± .013	.9211	+80%
	.02	$\geq$.981	.996 ± .003	.9278	+42%	1.000 ± .000	.9249	+92%
	.01	$\geq$.990	.998 ± .001	.9296	+45%	1.000 ± .000	.9276	+102%
Pilot floor	.5	=.531	.455 ± .099	.9021	+6%	.313 ± .103	.9035	+14%
	.3	=.718	.736 ± .082	.9069	+11%	.534 ± .100	.9070	+26%
	.2	=.811	.806 ± .080	.9100	+14%	.635 ± .092	.9092	+38%
	.1	=.902	.864 ± .072	.9143	+20%	.815 ± .080	.9135	+49%
	.05	=.956	.941 ± .044	.9193	+27%	.919 ± .049	.9171	+64%
	.02	=.984	.995 ± .005	.9236	+34%	.989 ± .011	.9208	+77%
	.01	=.991	1.000 ± .000	.9253	+37%	1.000 ± .000	.9228	+82%
Standard split conformal			.000	.8959	+0%	.000	.8988	+0%
Pilot only, conservative rank			1.000	.9130	+18%	1.000	.9110	+44%
Pilot only, plain rank			1.000	.8947	−1%	1.000	.8940	−2%

Shifted pilot: reading the table. The table illustrates Theorem 2(i), Proposition 1(i),(iv) and Theorem 3(i),(iv). Every rule that floors or rounds a local cutoff keeps its mean held-out coverage above nominal at every shift, as the marginal guarantee of upward rounding and flooring requires. The pilot-only threshold has no local guarantee, and its coverage falls to .855. The pilot floor loses identity first: its anchor level falls from .010 above the analysts' level at π_S=.5 to the level itself at π_S=.6 and to .012 below it at π_S=.7, that is by about five σ_F, and its identity falls from .898 to .388 and .025, as $H_n(F(T))^2$ predicts, while its coverage stays at that of the raw local cutoff (.9015) because the anchor no longer binds. The seeded coverage-unit rule's identity rises from .87 at π_S=0 to .93 at π_S=1, as the shifted-pilot extension of Theorem 2 (part (iii)(c) of its proof, Appendix C) predicts: its mismatch scales with the pilot's density relative to the analysts' at their quantile, which is 1.3 to 1.4 when the whole pilot comes from the large-score classes and .6 to .7 when it comes from S (`shift_diagnostics.py`). (At π_S=.5 its identity, .896 ± .012, is within one standard error of the certified .90.) The seeded score-unit rule's identity stays between .90 and .92, because its width depends on the pilot only through $\hat{f}$, the pilot's density at its own quantile, which moves from .83 at π_S=0 to .62 at π_S=.8 and back to .65 at π_S=1. The centred grid's identity is periodic in the gap, with dips to .79 and .75 at gaps of .41 and 1.49 cells and values of .95–1.00 at the positive gaps near whole cells, as the lattice law predicts, and it is less regular at negative gaps (.73 to .89, with .89 at −1.03 cells), where the lattice prediction fits less well (R^2 = .60). The coverage of both

Table S5: **Certificates across pilot sizes** (ρ=.1, n=15,000, 100 pilot redraws $\times$ 20 pairs per setting, since seedless identity varies mainly between pilots). Each rule is at its certified width or rank. For the centred grid the certificate is a lower bound, for the pilot floor the exact value, and for the seeded rule $1-\rho$. $\pm$ clustered by pilot.

		Centred grid				Pilot floor				Seeded, coverage units		
Pool	n_p	b	Certificate	Identity	Size	j	Exact	Identity	Size	b	Identity	Size
ImageNet	250	.0690	.914	$.976 \pm .011$	88.4	232	.907	$.899 \pm .029$	48.4	.0676	$.961 \pm .004$	63.4
	500	.0500	.904	$.971 \pm .013$	48.3	460	.907	$.884 \pm .031$	42.9	.0476	$.943 \pm .005$	48.3
	1,000	.0353	.904	$.948 \pm .017$	41.8	915	.915	$.902 \pm .027$	40.2	.0376	$.928 \pm .006$	44.0
	2,000	.0265	.903	$.991 \pm .005$	39.5	1,822	.912	$.918 \pm .025$	38.8	.0326	$.916 \pm .006$	42.3
	4,000	.0198	.900	$.955 \pm .016$	38.6	3,633	.907	$.906 \pm .027$	37.7	.0301	$.909 \pm .006$	41.4
Synthetic	250	.0690	.914	$.996 \pm .003$	64.8	232	.907	$.930 \pm .025$	22.6	.0676	$.963 \pm .004$	35.8
	500	.0500	.904	$.965 \pm .015$	22.1	460	.907	$.928 \pm .025$	19.9	.0476	$.945 \pm .005$	22.2
	1,000	.0353	.904	$.964 \pm .012$	18.8	915	.915	$.940 \pm .022$	18.1	.0376	$.936 \pm .005$	20.1
	2,000	.0265	.903	$.990 \pm .005$	18.0	1,822	.912	$.921 \pm .024$	17.1	.0326	$.922 \pm .005$	19.1
	4,000	.0198	.900	$.948 \pm .016$	16.9	3,633	.907	$.912 \pm .025$	16.6	.0301	$.910 \pm .006$	18.6

coverage-unit rules rises with the shift, from .913–.917 to .930–.931: above the analysts' quantile the shifted pilot has less density than the analysts' law, so one cell of the pilot's levels spans more of the analysts' coverage (at π_S=1, one cell above q spans about 1.7–1.8 cells of F). Upward rounding turns this into a size cost and keeps validity: mean set size grows from 39.3 to 50.6 for the centred grid and from 41.9 to 50.9 for seeded coverage-unit rounding, and stays at 42 for the score-unit rule, while the pilot-only threshold's shrinks from 38.6 to 22.7 as its coverage falls and the pilot floor's from 38.5 to 34.5, the size of the raw local cutoff.

Fresh-draw check. Table S7 removes the finite pool. Under every law the seeded rule's identity is .914 to .916 at ρ=.1, .957 to .958 at ρ=.05 and .991 to .992 at ρ=.01, which exceeds the certificate $1-\rho$ by .014 to .016, .007 to .008 and .0015: these are the values of $1/(n_p b)$, the pilot term $2\lambda/n_p$ of Theorem 2(iii-b) divided by the width, so the measured mismatch equals D_n/b, the bound for the oracle map, and the pilot term is slack on average. The measured identities agree across the five laws to within .003, as the law-free bound of Theorem 2(iii-b) leads one to expect. The centred grid exceeds its certificates by .04 to .05 at ρ=.1, .01 to .02 at ρ=.05 and at most .004 at ρ=.01, since the certificate counts the central cell only and the neighbouring cells contribute less as the width grows. The pilot floor's identity agrees with the exact value within 2.1 standard errors in all fifteen settings, on both sides (the largest deviations are $.9036 \pm .0042$ and $.9048 \pm .0042$ against .9122 at ρ=.1, and $.9914 \pm .0013$ against .9900 at ρ=.01), as an identity that holds for every continuous law must. Table S8 repeats the check at nine (n, n_p) pairs, including those of the language-model, CIFAR-100, ImageNet and census experiments, at $\alpha \in \{.05, .1, .2\}$, and for five and ten analysts. In all 39 settings the centred grid is at or above its certificate and the floor is within three standard errors of its exact value (38 of 39 within two). The seeded identity is at least $1-\rho$ in the 33 settings with $M = 2$ and at least $1-(M-1)\rho$, the union bound of Remark 3, for $M \in \{5, 10\}$. The mean coverages under fresh draws match the predictions of Section 7 to within .0005: at (15,000, 2,000, .1) the centred grid covers .9131/.9150/.9194 against the predicted $1-\alpha+b/2$ = .9133/.9153/.9198, and the floor .9106/.9125/.9169 against the exact $\mathbb{E}\max\{U, V_j\}$ = .9107/.9126/.9171, for ρ = .1/.05/.01. On the finite pools the held-out coverage of the floor and the grid is within .005 of these predictions at every ρ on all five pool settings (within .001 on ImageNet with APS), inside the sampling error of the fixed test split. The measured inflations subtract the standard cutoff's coverage on the same split, which is .9018 on ImageNet with APS against the expected .9000 and .8959 to .8988 on the other settings, and the split's error on the difference between two levels does not cancel, so the inflations are 83 to 110 percent of the predictions at $\rho \leq .1$ (88 to 91 percent on ImageNet with APS) and 59 to 120 percent at $\rho \geq .2$, where the predictions are at most .02 (Tables S2, S3 and S4). The floor's identity lies below its exact value in 24 of the 35 pool settings, by up to 2.1 clustered standard errors. A Gaussian model of the finite pool, with the reduced spread and the negative dependence of the two cutoffs (`finite_pool_model.py`), raises the exact values by at most .015, so this tendency is not a finite-pool effect that the model captures, and with 20 pilot redraws per setting, which share pilots across the two scores of a pool, it is within the pilot-to-pilot variability. Under fresh i.i.d. draws at (15,000, 2,000, .1) and ρ = .5/.3/.2 (5,000 draws, `v8_fresh_large_rho.py`), the floor's identity is .517/.720/.824 against the exact

Table S6: **A shifted pilot** (real ImageNet, ρ=.1, 40 pilot redraws $\times$ 25 pairs per row, the same analyst pairs in every row, details in the text). Rows with $\pi_S < .5$ over-represent the classes with large scores (negative gap), rows with $\pi_S > .5$ the classes with small scores (positive gap). π_S is the fraction of pilot images drawn from S, the 500 classes whose within-class 0.9-quantile of the APS score is below the median over classes (π_S=.5 is close to the analysts' mix). "Gap" is the pilot's level gap $G(q) - (1-\alpha)$ at the analysts' population quantile, in cells of the centred grid. Coverage on the held-out test split (nominal .90). "Pilot only" is the pilot threshold at the conservative rank, which uses the 2,000 pilot labels only, while the local rules also use each analyst's 15,000. For the pilot-only threshold, "undercov." is the fraction of pilot draws for which its held-out coverage is below .90. "Anchor" is the pilot floor's anchor level $F(Z_{(1822)}) - (1-\alpha)$ under the analysts' population law, whose sign decides the floor's identity by Proposition 1(iv) (it is $-.0001$ at π_S=.6). Top: held-out coverage and set size. Bottom: threshold identity $\pm$ its clustered standard error, integrated over the offset for the seeded rules.

		Pilot only		Pilot floor			Centred grid		Seeded, cov. units	Seeded, score units
π_S	Gap	Cov. (undercov.)	Size	Anchor	Cov.	Size	Cov.	Size	Cov.	Cov.
0	-1.76	.9404 (.00)	57.4	$+.040$	.9400	57.1	.9090	37.6	.9123	.9141
.1	-1.40	.9360 (.00)	53.8	$+.035$	.9355	53.5	.9114	38.8	.9130	.9147
.2	-1.03	.9308 (.00)	50.1	$+.030$	.9305	49.9	.9121	39.0	.9139	.9152
.3	-0.69	.9254 (.00)	46.5	$+.024$	.9249	46.2	.9110	38.6	.9147	.9159
.4	-0.32	.9184 (.00)	42.3	$+.017$	.9179	42.1	.9146	40.6	.9157	.9165
.5	0.03	.9113 (.00)	38.6	$+.010$	.9110	38.5	.9130	39.3	.9170	.9172
.6	0.41	.9022 (.30)	34.8	.000	.9042	35.4	.9139	40.4	.9186	.9177
.7	0.79	.8903 (.90)	31.0	$-.012$	.9015	34.5	.9194	43.3	.9207	.9181
.8	1.14	.8776 (1.00)	27.6	$-.024$	.9015	34.5	.9151	40.6	.9229	.9182
.9	1.49	.8660 (1.00)	24.9	$-.035$	.9015	34.5	.9218	45.3	.9261	.9178
1.0	1.82	.8545 (1.00)	22.7	$-.046$	.9015	34.5	.9313	50.6	.9297	.9174

		Threshold identity			
π_S	Gap	Pilot floor	Centred grid	Seeded, coverage units	Seeded, score units
0	-1.76	$1.000 \pm .000$	$.846 \pm .045$	$.869 \pm .014$	$.899 \pm .010$
.1	-1.40	$1.000 \pm .000$	$.734 \pm .054$	$.876 \pm .013$	$.904 \pm .010$
.2	-1.03	$1.000 \pm .000$	$.892 \pm .043$	$.882 \pm .013$	$.907 \pm .010$
.3	-0.69	$1.000 \pm .000$	$.852 \pm .048$	$.885 \pm .014$	$.912 \pm .009$
.4	-0.32	$1.000 \pm .000$	$.826 \pm .049$	$.892 \pm .014$	$.915 \pm .009$
.5	0.03	$.898 \pm .047$	$.951 \pm .031$	$.896 \pm .012$	$.919 \pm .008$
.6	0.41	$.388 \pm .073$	$.793 \pm .053$	$.901 \pm .012$	$.921 \pm .008$
.7	0.79	$.025 \pm .025$	$.897 \pm .038$	$.910 \pm .011$	$.924 \pm .008$
.8	1.14	$.000 \pm .000$	$.998 \pm .002$	$.917 \pm .009$	$.924 \pm .008$
.9	1.49	$.000 \pm .000$	$.754 \pm .056$	$.929 \pm .007$	$.921 \pm .009$
1.0	1.82	$.000 \pm .000$	$.985 \pm .015$	$.933 \pm .007$	$.919 \pm .009$

.516/.708/.819 (standard errors .007/.006/.005), the grid's is .790/.852/.896 against its certificates .513/.707/.807, and the seeded rule's is .604/.751/.833 against $1-\rho$.

F.3 Information regimes and competitors

Competitors. *Replicable binary search* adapts the replicable quantile search of Kalavasis et al. (2024, Theorem B.6) to the conformal level. It uses R=4,096 grid points over the score range and 12 rounds, and each round compares the empirical distribution function at the midpoint with a shared randomized target. Its guarantee is approximate two-sided discrete-quantile accuracy (Kalavasis et al., 2024, Theorem B.6), which is weaker than exact marginal validity. A realized coverage below $1-\alpha$ does not by itself contradict marginal validity, and the realized coverage of standard split conformal also falls below nominal on about half of its calibrations. We therefore report three versions. The symmetric target, uniform on $[1-\alpha-\varepsilon/2, 1-\alpha+\varepsilon/2]$, keeps mean coverage at the standard level (.899–.903). The upward-shifted target, uniform on $[1-\alpha, 1-\alpha+\varepsilon]$, has higher coverage, but it raises the acceptance level only by $\varepsilon/2$ on average and guarantees only $F(\tau) \geq 1-\alpha-\omega$ on a DKW event of radius ω. The version floored at the analyst's own conformal cutoff, $\max\{\text{search}, \widetilde{\tau}\}$, uses the upward-shifted target, so that the floor rarely binds, and is exactly marginally valid for every realization (Table S2, identity .39/.64/.80 on ImageNet at +11/38/107%). With the symmetric

Table S7: **Fresh-draw check of the certificates** (n=15,000, n_p=2,000, α=.1, 5,000 independent replications per law, each drawing a fresh pilot and two fresh calibration samples i.i.d. from the law, so that the assumptions of Theorem 2(iii-b), Theorem 3(ii) and Proposition 1(ii) hold exactly). Threshold identity, mean $\pm$ standard error over replications, of seeded coverage-unit rounding at $b = (D_n + 1/n_p)/\rho$ (identity integrated over the offset, certified $\geq 1 - \rho$), the centred grid at its smallest certified width (certified $\geq .9025/.9522/.9911$) and the pilot floor at its certified rank (exact $.9122/.9507/.9900$), for $\rho = .1/.05/.01$. Standard split conformal has identity 0 in every setting. The mixture is $.3\,\mathrm{LN}(0,1) + .7\,\mathrm{LN}(2,.5)$. `fresh_draw_check.py`.

Law	ρ=.1	ρ=.05	ρ=.01
Seeded, coverage units			
Uniform$(0,1)$	$.916 \pm .001$	$.958 \pm .001$	$.992 \pm .000$
Normal$(0,1)$	$.914 \pm .001$	$.957 \pm .001$	$.991 \pm .000$
Exponential(1)	$.916 \pm .001$	$.958 \pm .001$	$.992 \pm .000$
Beta$(2,5)$	$.915 \pm .001$	$.958 \pm .001$	$.992 \pm .000$
Lognormal mixture	$.914 \pm .001$	$.957 \pm .001$	$.991 \pm .000$
Centred grid			
Uniform$(0,1)$	$.950 \pm .003$	$.974 \pm .002$	$.994 \pm .001$
Normal$(0,1)$	$.939 \pm .003$	$.966 \pm .003$	$.992 \pm .001$
Exponential(1)	$.943 \pm .003$	$.969 \pm .002$	$.994 \pm .001$
Beta$(2,5)$	$.943 \pm .003$	$.968 \pm .003$	$.994 \pm .001$
Lognormal mixture	$.944 \pm .003$	$.971 \pm .002$	$.994 \pm .001$
Pilot floor			
Uniform$(0,1)$	$.910 \pm .004$	$.950 \pm .003$	$.990 \pm .001$
Normal$(0,1)$	$.914 \pm .004$	$.952 \pm .003$	$.990 \pm .001$
Exponential(1)	$.906 \pm .004$	$.946 \pm .003$	$.990 \pm .001$
Beta$(2,5)$	$.904 \pm .004$	$.947 \pm .003$	$.991 \pm .001$
Lognormal mixture	$.905 \pm .004$	$.950 \pm .003$	$.991 \pm .001$

Table S8: **Fresh-draw check at the sample sizes of the pilot-based experiments, at three levels, and for M analysts** (5,000 independent replications each, uniform scores, since the certificates do not depend on the law and Table S7 shows the measured identities agreeing across laws). Each replication draws a fresh pilot of n_p scores and M fresh calibration samples of n scores, and records whether all M analysts deploy the same threshold: seeded coverage-unit rounding at $b = (D_n + 1/n_p)/\rho$ (integrated over the offset, certified $\geq 1 - \rho$ for $M = 2$), the centred grid at its smallest certified width and the pilot floor at its certified rank, with the certificate for M analysts in brackets, $\mathbb{E}[(H_n(V_c) - H_n(V_a))_+^M]$ and the exact $\mathbb{E}[H_n(V_j)^M]$ (Remark 3). Standard errors are at most .006. The rows are the sample sizes of the CIFAR-100 released-probability pool (2,460), the language-model pools (2,625 and 13,500), the fine-tuned CIFAR-100 runs (4,921), Llama-3.2-1B (6,000), ImageNet (15,000), Folktables (20,000) and two smaller sizes. `v8_extras.py`.

				$\rho = .1$			$\rho = .05$			$\rho = .01$		
n	n_p	α	M	Seeded	Centred [cert]	Floor [exact]	Seeded	Centred [cert]	Floor [exact]	Seeded	Centred [cert]	Floor [exact]
500	250	.1	2	.922	.934 [≥.904]	.928 [=.932]	.961	.970 [≥.955]	.955 [=.956]	.992	.993 [≥.991]	.991 [=.992]
1,000	500	.1	2	.915	.920 [≥.902]	.908 [=.906]	.957	.963 [≥.950]	.961 [=.960]	.991	.992 [≥.992]	.989 [=.991]
2,460	937	.1	2	.916	.937 [≥.907]	.899 [=.902]	.958	.964 [≥.951]	.954 [=.956]	.992	.992 [≥.990]	.990 [=.991]
2,625	1,000	.1	2	.910	.934 [≥.904]	.911 [=.912]	.955	.964 [≥.954]	.949 [=.951]	.991	.990 [≥.990]	.990 [=.991]
4,921	1,875	.1	2	.910	.927 [≥.907]	.903 [=.901]	.955	.965 [≥.952]	.947 [=.950]	.991	.993 [≥.990]	.992 [=.991]
6,000	2,000	.1	2	.912	.936 [≥.902]	.910 [=.904]	.956	.970 [≥.953]	.955 [=.951]	.991	.992 [≥.990]	.993 [=.991]
13,500	2,000	.1	2	.915	.939 [≥.906]	.909 [=.909]	.958	.972 [≥.954]	.953 [=.956]	.992	.992 [≥.991]	.990 [=.991]
15,000	2,000	.1	2	.914	.944 [≥.903]	.907 [=.912]	.957	.969 [≥.952]	.945 [=.951]	.991	.993 [≥.991]	.989 [=.990]
20,000	2,000	.1	2	.918	.949 [≥.904]	.905 [=.907]	.959	.974 [≥.953]	.953 [=.955]	.992	.994 [≥.991]	.990 [=.991]
15,000	2,000	.05	2	.921	.953 [≥.915]	.911 [=.912]	.960	.973 [≥.956]	.953 [=.952]	.992	.992 [≥.991]	.992 [=.992]
15,000	2,000	.20	2	.913	.947 [≥.901]	.898 [=.900]	.956	.968 [≥.951]	.952 [=.952]	.991	.993 [≥.990]	.992 [=.991]
15,000	2,000	.1	5	.825	.893 [≥.858]	.878 [=.878]	.912	.944 [≥.927]	.934 [=.929]	.982	.989 [≥.985]	.988 [=.984]
15,000	2,000	.1	10	.769	.851 [≥.822]	.847 [=.851]	.885	.911 [≥.906]	.908 [=.911]	.977	.981 [≥.980]	.978 [=.979]

target and the floor, identity is .32 at $+5\%$ for $\varepsilon = .05$ on ImageNet. Theorem B.6 of Kalavasis et al. (2024) states its guarantee for $8\log_2^2(R)\ln(2/\beta)/(\varepsilon^2\rho^2)$ samples, above 10^7 at $R = 4{,}096$, $\beta = .1$ and every (ε, ρ) used here, so at $n = 15{,}000$ the search runs outside its guarantee and we report what it achieves. The *differentially private quantile* is the exponential mechanism over the grid with utility $-|\mathrm{rank} - k|$. Coupled through one shared uniform by inverse-CDF sampling, it has identity .022 at every ϵ_{DP}. Coupled by correlated sampling (Holenstein, 2007), the construction behind

Table S9: Full regimes table (master protocol, ρ=.1, coverage on the held-out test split). Rows with $\pm$ give the standard error clustered by pilot redraw. Rows without $\pm$ have identity fixed by construction or, for binary search and the DP quantile, binomial standard errors over pairs of at most .016 (.035 for correlated sampling, on 200 pairs). *Broadcast one threshold*: one analyst calibrates and sends its threshold to the other. *Pilot only*: the pilot order statistic at the plain rank $\lceil (1-\alpha)(n_p+1) \rceil$ or at the conservative rank, with sizes from the first four pilots (plain rank) or from all 20 (conservative rank, as in Table S1, while Table 2 uses all 1,000 pairs). *Pilot-centred* and *pilot-anchored grids*: seedless score-unit grids of plug-in width with a cell centred at the pilot's $(1-\alpha)$-quantile or a grid point at it. "Size-matched standard" is an oracle diagnostic, standard split conformal at the threshold whose *test* size equals RECAL's, and is not a deployable procedure. Binary search: symmetric target $1-\alpha+u_t$, or the upward-shifted target $1-\alpha+\varepsilon/2+u_t$, with $u_t \sim \mathrm{Unif}[-\varepsilon/2, \varepsilon/2]$. Neither is exactly marginally valid, whereas the floored version of Table S2 is. The pilot-floor row comes from `v6_regimes.py`, with identity and coverage from 1,000 pairs and size from 200 pairs over all 20 redraws. DP rows: the exponential mechanism over the grid with inverse-CDF coupling or with correlated sampling (Holenstein, 2007) through one shared stream (correlated-sampling rows on the 200 pairs of the first four redraws). The generic-width grid rounds in score units at the plug-in width with the Beta-law variance $\alpha(1-\alpha)/n$ replaced by $\log(8/\rho)/n$, twice the squared DKW radius of each run at level $\rho/4$, which is $\sqrt{\log(8/\rho)/(\alpha(1-\alpha))} = 6.98$ times the plug-in width at ρ=.1. Continuous-output alternatives are listed for coverage and size.

	Synthetic			Real ImageNet		
Method	Ident.	Coverage	Size	Ident.	Coverage	Size
Standard, local	.000 ± .000	.8988	14.8	.000 ± .000	.9018	34.6
Broadcast one threshold	1.000	.8988	14.8	1.000	.9018	34.6
Pooled	1.000 ± .000	.8990	14.9	1.000 ± .000	.9018	34.6
Pilot only, plain rank	1.000	.8951	14.1	1.000	.9015	34.7
Pilot only, conservative rank	1.000	.9077	16.9	1.000	.9117	38.8
Pilot-centred grid, score units (Prop. 7)	.977 ± .023	.9137	18.0	.999 ± .001	.9172	41.6
Pilot-anchored grid, score units	.952 ± .019	.9175	20.5	.831 ± .062	.9213	43.1
Centred grid, coverage units, b_0 (Thm 3)	1.000 ± .000	.9146	17.9	1.000 ± .000	.9174	41.3
Pilot floor, rank 1,822 (Prop. 1)	.846 ± .076	.9077	16.9	.949 ± .048	.9113	38.6
Seeded, coverage units, b_0 (Thm 2)	.940 ± .008	.9176	19.8	.911 ± .013	.9184	41.8
RECAL, plug-in width	.934 ± .008	.9153	18.9	.925 ± .012	.9176	41.8
RECAL, $\hat{\kappa}$=1.5	.956 ± .005	.9223	21.1	.950 ± .008	.9252	46.5
RECAL, conservative width	.974 ± .003	.9339	25.3	.978 ± .003	.9528	80.9
Generic-width grid (6.98× plug-in)	.991 ± .001	.9615	58.6	.989 ± .002	.9766	502.2
Size-matched standard (oracle diagnostic)	.000	.9159	18.9	.000	.9174	41.8
Bagged quantile (B=10)	.000	.8987	14.8	.000	.9020	34.6
Five-fold median quantile	.000	.8988	14.8	.000	.9019	34.6
Kernel-smoothed quantile (h=.02)	.000	.9009	15.2	.000	.9027	34.9
Binary search, symmetric target, ε=.02	.507	.8988	14.8	.405	.9023	34.9
Binary search, symmetric target, ε=.05	.672	.8989	15.5	.601	.9024	36.3
Binary search, symmetric target, ε=.1	.803	.9035	15.7	.752	.9007	35.5
Binary search, upward-shifted target, ε=.02	.487	.9088	17.4	.391	.9110	38.6
Binary search, upward-shifted target, ε=.05	.676	.9235	21.3	.613	.9261	46.8
Binary search, upward-shifted target, ε=.1	.839	.9527	46.8	.807	.9502	71.2
DP quantile, inverse-CDF coupling, ϵ_{DP}=10	.022	.8988	14.8	.022	.9019	34.6
DP quantile, correlated sampling, ϵ_{DP}=.05	.595	.8987	14.7	.445	.9022	34.6
DP quantile, correlated sampling, ϵ_{DP}=.2	.215	.8985	14.8	.125	.9020	34.7
DP quantile, correlated sampling, ϵ_{DP}=1	.035	.8986	14.8	.060	.9018	34.6

the privacy-to-replicability conversion of Bun et al. (2023), it reaches identity $.595/.215/.035$ at $\epsilon_{\mathrm{DP}} = .05/.2/1$ on the synthetic pool ($.445/.125/.060$ on ImageNet). The mechanism is symmetric about the target, so its coverage stays at the standard level with no guarantee attached, and higher identity needs a smaller ϵ_{DP} and hence a noisier threshold. The *bagged*, *five-fold-median* and *kernel-smoothed* quantiles are scalar stabilization baselines implemented for this experiment. They are not implementations of the full methods of Ndiaye (2022) or Gupta et al. (2022). Their outputs have atomless laws conditionally on the shared randomness, so their identity is 0 by Lemma 1, and the tables record this value without measuring it. *Pilot-only*, *pooled* and *broadcast* rules have identity 1 by construction. The coverage of the pilot-only threshold at the plain rank is one $\mathrm{Beta}(1801, 200)$ draw (standard deviation .0067). Under the master protocol it lies in $[.89, .92]$ for 15 of the 20 synthetic pilots, and for 19 of 20 at the conservative rank. Over 200 pilot redraws of the in-pool protocol, the plain-rank coverage has minimum .882 and falls below nominal for 52% of pilots.

Pilot-centred grid in score units. Averaged over 40 pilot redraws of each size, each re-splitting the pool, the score-unit pilot-centred grid (Proposition 7) has identity $0.917/0.953/0.995/0.974/1.000$ with pilots of $100/300/1{,}000/2{,}000/3{,}000$ points (RECAL's random offset on the same pairs: 0.914–0.917). With one pilot and one split of the synthetic pool, n=15,000, the plug-in width at ρ=.1 and 300 draws, all M=2/5/10/20 analysts coincide with probability 1.00 (RECAL: $0.96/0.91/0.88/0.85$). In this run the M analysts alternate between the two halves, so for $M \geq 3$ two analysts drawing from the same half overlap in most of their 15,000 points, which makes them more alike than independent analysts: with M independent in-pool analysts at the same width, all coincide under RECAL with probability .845/.818/.763 for $M = 5/10/20$ (`v5_experiments.py`, key `X4_multi_analyst`), against the range-law prediction .856/.809/.768 of Proposition 4(iii), and the fresh-draw rows for $M \in \{5, 10\}$ in Table S8 are the clean test for the coverage-unit rules. The single split explains why .96 exceeds the master mean .934. At n_p=100 the density estimate uses the first 2,000 points of the redraw. Averaged over 20 pilots, a location shift of analyst B by $0/.1/.25/.5/1$ cells gives $0.98/0.99/0.82/0.39/0.02$ for the pilot-centred grid and $0.93/0.89/0.74/0.49/0.03$ for RECAL. Both fall steeply near half a cell, as Proposition 4(iv) and Proposition 7 describe for an analyst-side shift.

Cell-mass diagnostic. The premium is driven by the label mass of the part of a grid cell that upward rounding adds. With the mean label count $\mathcal{H}(t) := \mathbb{E}_X \#\{y : s(X, y) \leq t\}$ computed on unlabeled test inputs and a quantile estimate $\hat{q}$, the integrated diagnostic $\int_0^1 \mathcal{H}(\hat{q} + u\beta)\, du/\mathcal{H}(\hat{q}) - 1$ is the expected relative size increase of a uniform upward excess started at $\hat{q}$. The width β comes from the pilot and $\hat{q}$ is the conformal quantile of the two pooled analyst halves (38,000 points), so the diagnostic approximates the premium of a random calibration threshold near the pool quantile. It gives 23% for synthetic APS (measured +27% on 52 pairs, +28% on all 1,000), 25% for synthetic RAPS (+28%), 20% for ImageNet APS (+21% and +22%) and 15% for ImageNet RAPS (+15% and +13%). The linearized version $\{\mathcal{H}(\hat{q}+\beta) - \mathcal{H}(\hat{q})\}/(2\mathcal{H}(\hat{q}))$ gives 24/25/21/18% on the image pools, within 3.1 points of the integrated one, but overshoots on the language models, where $\mathcal{H}$ is convex over a cell (691% against a measured +143% for GPT-2 APS).

F.4 FURTHER ARCHITECTURES, LANGUAGE MODELS, AND SITES

Table S10: Two further architectures under the master protocol: CIFAR-100 logits of a fine-tuned ResNet-18 (held-out top-1 .809) and ViT-B/16 (.876), N=20,000, K=100. Test 5,000, pilot 1,875, disjoint halves of 6,562 (one point unused), n=4,921 per analyst (75% of a half), ρ=.1, 1,000 pairs over 20 pilots. These runs, made with an earlier version of the master-protocol script (Reproducibility map), have no coverage-unit rows and no exact offset integration, so seeded sizes are four-offset estimates and seeded coverages sixteen-offset averages (see Estimators). Premiums exceed ImageNet's because n is a third as large ($\beta \propto n^{-1/2}$).

	ResNet-18				ViT-B/16			
Method	Ident.	Coverage	Size	Premium	Ident.	Coverage	Size	Premium
Standard APS	.000	.9017	5.52	+0%	.000	.9075	3.17	+0%
Pilot only, conservative rank	1.000	.9180	6.30	+14%	1.000	.9197	3.61	+14%
Pilot-centred grid, score units	1.000	.9330	7.52	+36%	.958	.9343	4.26	+35%
RECAL, plug-in width	.923	.9301	7.70	+40%	.895	.9330	4.25	+34%
RECAL, $\hat{\kappa}$=1.5	.949	.9414	9.82	+78%	.930	.9453	5.39	+70%
RECAL, conservative width	.978	.9712	40.02	+625%	.969	.9739	26.24	+729%
Standard RAPS	.000	.9023	3.06	+0%	.000	.9074	1.88	+0%
RECAL-RAPS, plug-in width	.923	.9279	3.58	+17%	.892	.9319	2.18	+16%

Two further architectures. On CIFAR-100 (Table S10), with a fine-tuned ResNet-18 (He et al., 2016) and ViT-B/16 (Dosovitskiy, 2020), RECAL reaches identity 0.923 (ResNet-18) and 0.895 (ViT-B/16) at +40% and +34%. The ViT values are below the 0.900 target, by .005 (APS) and .008 (RAPS), with standard errors over pairs of about .003 (clustered errors were not recorded for these runs), consistent with the plug-in width, whose calibration is asymptotic. The networks are torchvision `ResNet18_Weights.IMAGENET1K_V1` and `ViT_B_16_Weights.IMAGENET1K_V1`, fine-tuned on CIFAR-100. The score-unit pilot-centred grid reaches 1.000 and 0.958 at +36% and +35%, and RECAL-RAPS reaches 0.923 and 0.892 at +17% and +16%. The conservative width

costs $+625\%$ and $+729\%$ at this n. At $(n, n_p) = (4{,}921, 1{,}875)$, coverage-unit rounding at b_0 would be certified at mismatch .089, and the centred grid at identity .9995.

Language models. Logits were produced by `extract_logits.py` (one GPU, `float16`) for the Hugging Face checkpoints `gpt2` (Radford et al., 2019), `EleutherAI/pythia-1.4b` (Biderman et al., 2023) (and `-70m`, `-160m`, `-410m`, `-1b` for the sweep), `Qwen/Qwen2.5-1.5B` (Qwen et al., 2025) and `meta-llama/Llama-3.2-1B` (Grattafiori et al., 2024), on test contexts of length 512 from `wikitext-103-raw-v1` (Merity et al., 2016). Model revisions were not pinned. Per model (N, K, top-1/top-5, mean entropy): GPT-2 (10,000, 50,257, .402/.624, 3.59 nats), Pythia-1.4B (10,000, 50,304, .504/.729, 2.46), Qwen2.5-1.5B (10,000, 151,936, .519/.754, 2.18) and Llama-3.2-1B (20,000, 128,256, with .516/.751 and 2.31 measured on a 5,000-context subset). Contexts come from one test split, and overlapping windows from the same document make the pool dependent at the document level. Table S11 runs the four models under the master protocol. Seeded score-unit rounding at the plug-in width has mean identity of at least $1 - \rho$ in all 42 settings of model, score, n and $\rho \in \{.1, .05, .01\}$ in Tables S11 and S12, with mismatch between $.49\rho$ and $.86\rho$, against the $.80\rho$ that Proposition 2 predicts for an exact density estimate. At ρ=.1, $\hat{\kappa}$=1, APS identity is $0.939/0.920/0.914/0.926$ (GPT-2/Pythia/Qwen/Llama) and RAPS identity $0.942/0.944/0.951/0.942$. Standard calibration has identity 0.000 on every model. The coverage inflation at ρ=.1, measured as ReCal coverage minus standard coverage on the same test split, is .045/.037/.040/.030 for APS and .031/.026/.033/.025 for RAPS (GPT-2/Pythia/Qwen/Llama). For the plug-in width, Proposition 2(ii) predicts $\sqrt{2\alpha(1-\alpha)/n}/(2\rho)$ when the pilot's density estimate is exact, which is .041 at n=2,625 and .027 at n=6,000. The identities are means over ten pilot redraws. The per-redraw values were not stored, so the standard errors over pairs do not include the variation between redraws. Under the same design on ImageNet, standard errors clustered by redraw are about .01 at ρ=.1 (Table 2). The relative mean premium of RAPS is lower than that of APS on GPT-2, Pythia and Qwen ($+128/130/137\%$ against $+143/165/183\%$) and higher on Llama ($+102\%$ against $+70\%$, at n=6,000). The two scores differ mainly in the tail of the set-size distribution. At ρ=.05 the APS map's mean set size exceeds half the vocabulary in 34–41% of pair–offset draws on GPT-2, Pythia and Qwen and 8% on Llama, and at ρ=.01 in 80–93%. RAPS never does, and its 99th-percentile mean size is at most 1,292 tokens at ρ=.05. The integrated cell-mass diagnostic predicts the RAPS premium within 15 points on three models and 30 on GPT-2 ($98/116/145/101\%$ predicted against $+128/130/137/102\%$ measured). It predicts the APS premium within a factor 1.7 ($236/147/167/63\%$ against $+143/165/183/70\%$). The 300-pair spectrum streams with sampled offsets, drawn from one pilot split, give APS $0.973/0.960/0.950/0.867$ and RAPS $0.953/0.960/0.950/0.987$. The between-split component, whose per-redraw values were not stored, may explain why a single-split stream can be up to .06 below the multi-split mean. The seedless rules were not run on the language-model pools, whose logits are not in the archive, and neither was seeded rounding in coverage units. Their certificates at the language-model sizes follow from `certificates.py`: at $(n, n_p) = (2{,}625, 1{,}000)$ the centred grid is certified at identity .904/.954/.990 at widths .043/.049/.063 in coverage units, the pilot floor at ranks $918/921/928$ has exact identity .912/.951/.991, and seeded coverage-unit rounding needs widths .076/.152/.760, for ρ = .1/.05/.01. At (13,500, 2,000) the grid's widths are .027/.032/.040 (certificates .906/.954/.991), the floor's ranks 1,822/1,827/1,836 (exact .909/.956/.991), and the seeded widths .034/.068/.341. The premiums of these rules on the language-model pools are not measured. The comparison of n=2,625 with n=13,500 in Tables S11 and S12 changes the pool and the pilot along with n, so it is not a pure calibration-size effect.

Language models at n=13,500. Table S12 reruns GPT-2, Pythia-1.4B and Qwen2.5-1.5B on 40,000-context pools. At ρ=.1, $\hat{\kappa}$=1, the APS premium falls from $+143/165/183\%$ at n=2,625 to $+35/33/35\%$ at identity $0.933/0.915/0.917$. This is faster than the factor 2.3 implied by $\beta \propto n^{-1/2}$, because the label-count curve is convex. The RAPS premium falls to $+59/49/70\%$ at identity $0.951/0.942/0.950$, now above APS's in relative terms because its standard sets are 2–3 times smaller. APS still has the heavier tail. At ρ=.05 and $\hat{\kappa}$=1 no APS draw exceeds half the vocabulary any more (99th percentiles $1{,}238/360/473$ tokens). The fraction exceeding half the vocabulary is $5/10/10\%$ (GPT-2/Pythia/Qwen) at $\hat{\kappa}$=1.5 and ρ=.05, and at ρ=.01 it is 72–75% at $\hat{\kappa}$=1 and 77–83% at $\hat{\kappa}$=1.5. RAPS never crosses half the vocabulary at any ρ (99th percentile at most 2,376 tokens at ρ=.01). The coverage inflation at ρ=.1 is .018/.014/.018 for APS and .017/.014/.021 for RAPS, against the predicted .018. The single-split 300-pair streams of this run (.83/.87/.78 APS,

Table S11: Language models under the master protocol (n=2,625 for GPT-2, Pythia and Qwen and n=6,000 for Llama, splits in the protocol paragraph, 1,000 pairs over 10 pilot redraws, sizes on 40 pairs $\times$ 4 fixed offsets for RECAL and on 40 pairs for standard calibration, see Estimators). Identity: RECAL at $\hat{\kappa}$=1 | 1.5 (standard errors over pairs at most .002, without clustering by redraw). Coverage and mean set size: standard calibration $\rightarrow$ RECAL at $\hat{\kappa}$=1, with the premium and the 99th percentile of the mean set size over pair–offset draws. $> K/2$: fraction of pair–offset draws whose mean set size exceeds half the vocabulary, at $\hat{\kappa}$=1 | 1.5. The four-offset size estimates are biased where part of a cell deploys the full vocabulary (APS at $\rho \leq .05$, see Estimators), and the premiums there are indicative.

Score	ρ	Identity	Coverage	Mean size	Premium	99th pct.	$> K/2$
GPT-2 (50k vocabulary), n=2,625							
APS	.1	.939 | .959	.900 → .945	358 → 869	+143%	2,572	.00 | .12
APS	.05	.969 | .980	.900 → .971	358 → 18,486	+5,063%	50,257	.34 | .56
APS	.01	.994 | .996	.900 → .994	358 → 46,650	+12,929%	50,257	.93 | .98
RAPS	.1	.942 | .961	.907 → .938	193 → 441	+128%	797	.00 | .00
RAPS	.05	.971 | .981	.907 → .951	193 → 618	+220%	1,292	.00 | .00
RAPS	.01	.994 | .996	.907 → .981	193 → 2,756	+1,327%	5,404	.00 | .00
Pythia-1.4B (50k), n=2,625							
APS	.1	.920 | .947	.909 → .945	111 → 294	+165%	989	.00 | .22
APS	.05	.960 | .973	.909 → .972	111 → 20,985	+18,781%	50,304	.41 | .60
APS	.01	.992 | .995	.909 → .995	111 → 46,556	+41,789%	50,304	.93 | .88
RAPS	.1	.944 | .962	.907 → .933	46 → 105	+130%	183	.00 | .00
RAPS	.05	.972 | .981	.907 → .945	46 → 156	+242%	304	.00 | .00
RAPS	.01	.994 | .996	.907 → .975	46 → 759	+1,567%	1,578	.00 | .00
Qwen2.5-1.5B (152k), n=2,625							
APS	.1	.914 | .943	.912 → .952	99 → 282	+183%	1,989	.00 | .21
APS	.05	.957 | .971	.912 → .975	99 → 54,466	+54,712%	151,936	.36 | .68
APS	.01	.991 | .994	.912 → .995	99 → 133,200	+133,947%	151,936	.88 | .85
RAPS	.1	.951 | .967	.902 → .935	35 → 82	+137%	138	.00 | .00
RAPS	.05	.975 | .984	.902 → .949	35 → 124	+258%	231	.00 | .00
RAPS	.01	.995 | .997	.902 → .978	35 → 645	+1,761%	1,212	.00 | .00
Llama-3.2-1B (128k), n=6,000							
APS	.1	.926 | .950	.902 → .932	80 → 135	+70%	244	.00 | .00
APS	.05	.963 | .975	.902 → .957	80 → 10,813	+13,456%	128,256	.08 | .34
APS	.01	.993 | .995	.902 → .991	80 → 102,671	+128,620%	128,256	.80 | .90
RAPS	.1	.942 | .961	.903 → .928	30 → 60	+102%	102	.00 | .00
RAPS	.05	.971 | .981	.903 → .942	30 → 90	+205%	157	.00 | .00
RAPS	.01	.994 | .996	.903 → .970	30 → 373	+1,163%	768	.00 | .00

.96/.96/.95 RAPS) are up to .14 below the ten-split means, which the between-split component described above may explain.

Model-scale sweep (Pythia 70M–1.4B, in-pool protocol, fixed window). Five sizes of one family, sharing one tokenizer and one 50,304-token vocabulary over a 20$\times$ parameter range, isolate model quality (Table S13, Figure S1). The pilot density stays in $[0.96, 1.15]$, and the same grid rule with no per-model tuning gives identity .85–.94 at ρ=.1 and .94–.96 at ρ=.05. With $\hat{\kappa}$=1.5, mismatch is at or below the target except for Pythia-1.4B ($.103 \pm .018$). Across the sweep, standard mean set size falls from 559 to 109 tokens and pointwise agreement rises from .16 to .40. These pointwise agreements say nothing about exact identity, which is at most .01 for standard calibration (ties in the finite pool). The size cost varies non-monotonically within +46–84%.

Multi-site replication at scale. Table S14 calibrates one frozen income model, trained on Illinois, at pairs of five states with n up to 20,000 per site. Identity ranges from .585 (NY–PA, n=20,000, ρ=.1) to .995. The gap between the two sites' population quantiles, measured in grid cells, tracks it: NY–PA has a gap of 0.42 cells at n=20,000, while TX–PA has 0.03 cells and identity .910. A larger n shrinks the cell and turns a fixed quantile gap into a larger fraction of a cell. The gap is an empirical diagnostic. The location-shift bound of Proposition 4(iv) assumes that the two score laws differ only by a location shift, which distinct state populations need not satisfy. A distribution-free alternative is $\mathbb{P}(\tau_A \neq \tau_B) \leq \min\{1, (|q_A - q_B| + \mathbb{E}|\widetilde{\tau}_A - q_A| + \mathbb{E}|\widetilde{\tau}_B - q_B|)/\beta\}$. Each site keeps its own coverage throughout (Table S15).

Table S12: Language models at n=13,500 per analyst (40,000-context pools, pilot 2,000, disjoint halves of 18,000, held-out test 2,000, 1,000 pairs over 10 pilot redraws, sizes on 25 pairs × 4 fixed offsets for RECAL and on 25 pairs for standard calibration). Columns as in Table S11 (identity standard errors over pairs at most .0022, without clustering by redraw), with the same four-offset caveat for APS at $\rho \leq .05$.

Score	ρ	Identity	Coverage	Mean size	Premium	99th pct.	$> K/2$
GPT-2, n=13,500							
APS	.1	.933 | .955	.898 → .916	341 → 460	+35%	621	.00 | .00
APS	.05	.966 | .978	.898 → .938	341 → 670	+97%	1,238	.00 | .05
APS	.01	.993 | .996	.898 → .986	341 → 36,434	+10,598%	50,257	.72 | .83
RAPS	.1	.951 | .968	.901 → .918	171 → 272	+59%	399	.00 | .00
RAPS	.05	.976 | .984	.901 → .928	171 → 351	+105%	580	.00 | .00
RAPS	.01	.995 | .997	.901 → .961	171 → 1,227	+617%	2,376	.00 | .00
Pythia-1.4B, n=13,500							
APS	.1	.915 | .943	.910 → .925	100 → 133	+33%	180	.00 | .00
APS	.05	.957 | .972	.910 → .940	100 → 187	+87%	360	.00 | .10
APS	.01	.991 | .994	.910 → .986	100 → 37,798	+37,778%	50,304	.75 | .79
RAPS	.1	.942 | .961	.916 → .930	46 → 68	+49%	96	.00 | .00
RAPS	.05	.971 | .981	.916 → .940	46 → 94	+107%	146	.00 | .00
RAPS	.01	.994 | .996	.916 → .965	46 → 300	+559%	620	.00 | .00
Qwen2.5-1.5B, n=13,500							
APS	.1	.917 | .945	.902 → .921	84 → 113	+35%	154	.00 | .00
APS	.05	.958 | .972	.902 → .940	84 → 180	+115%	473	.00 | .10
APS	.01	.992 | .994	.902 → .986	84 → 110,983	+132,191%	151,936	.73 | .77
RAPS	.1	.950 | .967	.898 → .918	29 → 50	+70%	76	.00 | .00
RAPS	.05	.975 | .983	.898 → .930	29 → 70	+138%	116	.00 | .00
RAPS	.01	.995 | .997	.898 → .962	29 → 245	+731%	509	.00 | .00

Table S13: Pythia scale sweep (in-pool protocol, WikiText-103 pools of 10,000 contexts, n=5,000, ρ=0.1, 100 table pairs and 300 safety pairs per size, binomial standard errors). Standard: mean set size and pointwise agreement of standard calibration. RECAL at the plug-in width, and the last column at $\hat{\kappa}$=1.5 on the 300 safety pairs.

				Standard		RECAL		
Model	Top-1	Entropy	$\hat{f}$	Size	Agreement	Identity	Premium	Mismatch, $\hat{\kappa}$=1.5
Pythia-70m	.343	3.89	1.15	559	.16	.87	+46%	.070 ± .015
Pythia-160m	.412	3.53	1.00	342	.28	.85	+47%	.060 ± .014
Pythia-410m	.467	2.75	0.96	162	.32	.89	+65%	.070 ± .015
Pythia-1b	.489	2.54	0.98	127	.38	.94	+84%	.033 ± .010
Pythia-1.4B	.504	2.46	1.10	109	.40	.93	+58%	.103 ± .018

A pilot from a third site. Table S16 puts the analysts at two states and the pilot at a third. It evaluates the score-unit pilot-centred grid, and the centred grid in coverage units was not run on these sites. When settings are grouped by the pilot's gap from the analysts' mean quantile, the pilot-centred grid matches or exceeds RECAL on identity: 1.000 against .920 below a tenth of a cell, .974 against .863 up to a quarter cell, and a tie at .91–.92 beyond. Each site's own coverage stays above nominal. The pilot-only threshold keeps identity 1 but covers as little as .893 at the sites whose quantile it does not match. The bins group by the gap to the *mean* analyst quantile, which does not bound the gap to each analyst, and they cover only the gaps these five states produce. When all five states share one pilot at n=5,000 (20 pilot redraws × 10 trials for each pilot state, which is itself one of the five), the pilot-centred grid's thresholds coincide in every trial, and RECAL's offset-integrated all-five coincidence is 0.80 at ρ=.1 and 0.90 at ρ=.05 (averaged over the five pilot states).

Clinical sites (small n, in-pool coverage). The four UCI Heart Disease sites (Janosi et al., 1989) are the Cleveland Clinic (303 patients, of which 153 remain after the training split), the Hungarian Institute of Cardiology (294), V.A. Long Beach (200) and University Hospital Zurich (123). The data are licensed under CC BY 4.0. A logistic score on seven cross-site-complete features is fitted on a Cleveland split of 150 patients (.85 training accuracy) and frozen, with deterministic instance jitter and a pilot density from the training scores ($\hat{f}$=.67, window half-width .05). Analysts calibrate on their own hospital with unequal sizes, n_A=120 at Cleveland and n_B=200/197/120 at Budapest/Long

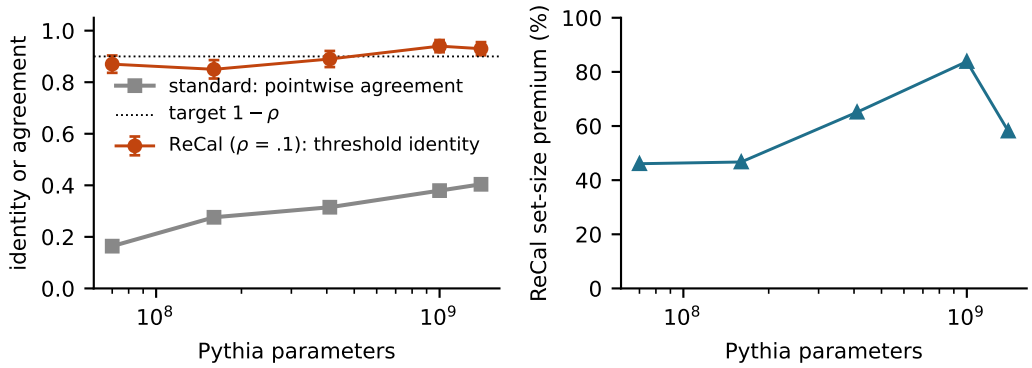


Figure S1: Replication cost against model scale for five Pythia models with one 50k-token vocabulary (in-pool protocol, ρ=.1, plug-in width, 100 pairs per model). Left: threshold identity of ReCal with binomial standard errors, pointwise agreement of standard split conformal, and the target $1-\rho$ (dotted). Three of the five identities (.87, .85, .89) lie below the target, within 1.4 standard errors of it, and the independent 300-pair safety streams give mismatch .103/.087/.087/.100/.097 at this width (standard errors about .017, so Pythia-1.4B's .097 and its .103 at $\hat{\kappa}$=1.5 in Table S13 agree within noise). Right: the set-size premium of ReCal.

Table S14: Multi-site replication at scale (Folktables `ACSIncome` (Ding et al., 2021), 2018 1-Year person files, five states as sites, N per site 68k–196k, a frozen logistic regression on standardized features trained on Illinois, shared pilot density $\hat{f}$=.66 from Illinois scores). In-pool protocol: each analyst resamples its state's pool with replacement, one shared offset per trial, coverage on the whole state pool. Width 1.5× the plug-in width, 200 trials per setting. Each entry: identity of the two sites' deployed thresholds (a multiple of 1/200), and in parentheses the gap between the sites' population quantiles in grid cells (an empirical diagnostic).

	n=1,000		n=5,000		n=20,000	
Site pair	ρ=.1	ρ=.05	ρ=.1	ρ=.05	ρ=.1	ρ=.05
CA-TX	.905 (.05)	.955 (.03)	.870 (.12)	.895 (.06)	.750 (.24)	.885 (.12)
CA-NY	.915 (.03)	.955 (.02)	.910 (.07)	.965 (.04)	.845 (.14)	.930 (.07)
CA-FL	.930 (.04)	.995 (.02)	.920 (.10)	.920 (.05)	.800 (.20)	.920 (.10)
CA-PA	.935 (.06)	.970 (.03)	.865 (.14)	.895 (.07)	.735 (.27)	.890 (.14)
TX-NY	.905 (.09)	.945 (.04)	.815 (.19)	.885 (.10)	.630 (.39)	.845 (.19)
TX-FL	.945 (.01)	.970 (.01)	.940 (.02)	.945 (.01)	.930 (.05)	.975 (.02)
TX-PA	.935 (.01)	.970 (.00)	.925 (.01)	.950 (.01)	.910 (.03)	.965 (.01)
NY-FL	.915 (.08)	.950 (.04)	.860 (.17)	.910 (.08)	.655 (.34)	.845 (.17)
NY-PA	.885 (.09)	.930 (.05)	.810 (.21)	.890 (.10)	.585 (.42)	.800 (.21)
FL-PA	.940 (.02)	.980 (.01)	.940 (.04)	.950 (.02)	.890 (.08)	.945 (.04)

Beach/Zurich (Corollary 4), with the width $2\hat{\kappa}B_n^+/(\hat{f}\rho)$ at $\hat{\kappa}$=1.5 and 400 trials, each with one sampled offset (binomial standard errors at most .021). Coverage is computed on each site's whole pool, including its calibration patients, and is therefore an in-pool quantity. At Long Beach and Zurich only three patients are left out of each calibration sample. Baseline identity is 0.00 for every pair. At ρ=.3 and .5, identity is .91/.84 (Cleveland–Budapest), .88/.78 (Cleveland–Long Beach) and .96/.94 (Cleveland–Zurich), with near-vacuous sets of 1.72–1.89 labels out of 2. At hospital scale, identity is practical only at coarse ρ. A simultaneous two-sample DKW bound on the score-CDF discrepancy of Corollary 5, $\eta_P \le \mathrm{KS} + \sqrt{\log(4/\delta)/2}\,(N_A^{-1/2} + N_B^{-1/2})$ with the site pool sizes, gives .28–.43 at δ=.05 from the measured KS distances of .08–.17. This is too loose to instantiate the corollary. No independent pilot exists at this scale ($303 - 150 - 120 = 33$ Cleveland patients remain), so the only pilot-only comparison available reuses the training split: a threshold calibrated on the 150 training patients covers 0.863 on the remaining Cleveland patients and 0.840/0.855/0.870 elsewhere. Those patients were also used to fit the model, so this comparison mixes training-data reuse with site shift and does not isolate the value of local calibration.

Regression. On UCI abalone (4,177 rows), ridge regression (penalty 1, one-hot sex, split seed 11) is fitted on 2,000 rows, and the absolute residuals on the remaining 2,177 form the score pool. We

Table S15: Folktables, own-site coverage of the deployed RECAL threshold at n=20,000 (site A / site B, in-pool protocol as in Table S14).

Site pair	ρ=.1	ρ=.05
CA-TX	.917 / .917	.931 / .932
CA-NY	.917 / .918	.931 / .933
CA-FL	.917 / .919	.931 / .935
CA-PA	.917 / .918	.931 / .934
TX-NY	.918 / .918	.934 / .933
TX-FL	.918 / .919	.934 / .935
TX-PA	.918 / .918	.934 / .934
NY-FL	.919 / .918	.934 / .935
NY-PA	.919 / .918	.934 / .934
FL-PA	.919 / .918	.936 / .934

Table S16: The shifted-pilot regime on census sites (Folktables `ACSIncome`, frozen Illinois model, analysts at two of CA/TX/NY/FL/PA sampling without replacement from their state pools, coverage on the whole state pool, the 2,000-point pilot from a third state, $n \in \{5{,}000, 20{,}000\}$, $\rho \in \{.1, .05\}$, $\hat{\kappa}$=1.5, 200 pairs over 20 pilot redraws per setting, 120 settings). Rows group settings by the pilot's quantile gap from the analysts' mean quantile, in grid cells. Identity of the score-unit pilot-centred grid and of RECAL, and the lowest own-site coverage across the two analysts for each rule (nominal .90).

		Identity		Lowest own-site coverage		
Pilot gap (cells)	Settings	Pilot-centred	RECAL	Pilot only	Pilot-centred	RECAL
$<$ .1	62	1.000	.920	.893	.930	.937
.1–.25	53	.974	.863	.893	.912	.920
.25–.5	5	.914	.918	.893	.912	.920

use deterministic jitter, a pilot density of $\hat{f}$=0.05 from the fixed window of half-width .02, which on the residual scale of several rings holds a single point of the 500-point pilot, so the width is set with large uncertainty, 300 pairs, and with-replacement sampling from the pool (in-pool protocol). At $(n, \rho) = (1000, .3)/(1000, .1)/(2000, .1)$, identity is $0.81/0.93/0.94$ (standard calibration: at most 0.013, from ties in the finite pool), coverage is $0.922/0.947/0.938$, and the interval half-widths are 3.87, 4.74 and 4.38 rings, against 3.43 (n=1,000) and 3.42 (n=2,000) for standard calibration.

Answer-set cache. Two serving sites calibrate the conformal-LM threshold of Quach et al. (2024) on disjoint question sets and cache the deployed answer sets. A cache entry is reusable across sites only if the two sets coincide. The pool is the union of the WebQuestions (Berant et al., 2013) train and test splits (`stanfordnlp/web_questions`) minus the 5 questions used as few-shot examples, which leaves 5,805 questions, each with K=20 sampled answers. The score is minus the best mean token log-probability among correct samples (token-F1 $\geq$ .5 against any gold alias), or 10^6 if no sample is correct. The pool is split into a pilot of 967 questions and two sites of 2,419. The width is the plug-in width with $\hat{\kappa}$=1.5. The level α is raised to $1 -$ (fraction of questions with a correct sample) $+$ .10 (.48 for Llama-3.2-1B, .41 for 3B), computed on the whole pool. Coverage, answer counts and cache hits are computed on all 5,805 questions, which include the calibration questions, so this is an in-pool evaluation. Identity is integrated exactly over the offset on 300 pairs. Cache hits use the same 300 pairs at 4 fixed offsets, and sizes and coverage use 20 of them. Fixed offsets estimate the offset average with a bias that depends on the positions of the thresholds relative to the grid. At n=2,419 each site's sample is its whole site, so all pairs coincide and the rows describe one site partition. The table illustrates the deployment effect and gives no held-out estimate.

Reading the cache table. In Table S17, under standard calibration 17–39% of cached answer sets differ between the sites. Under RECAL, 0–11% differ at the four evaluated offsets, at 1.7–2.4 times as many cached answers (four-offset estimates). Because identity lower-bounds pointwise agreement (Definition 1), the expected share of differing cached sets over the offset is at most one minus the identity, which is exact: 5.2% and 2.7% at n=2,419 and ρ=.1 for Llama-3.2-1B and 3B. The excess coverage of RECAL over standard calibration is 6.6 and 9.2 points for Llama-3.2-1B at n=2,419 and $\rho = .1, .05$. A cache hit printed as 1.00 (at least .995) means that almost no disagreement was observed at the evaluated offsets. It does not establish zero disagreement probability, because identity

Table S17: Answer-set cache across two serving sites (in-pool evaluation, details in the text). Answers cached: mean number of distinct answers in a deployed answer set. Cache hit: fraction of questions whose deployed answer sets are identical at the two sites (standard → RECAL, the latter at 4 fixed offsets). At n=2,419 all pairs coincide (one site partition).

n	ρ	Identity	Coverage std→RECAL	Answers cached	Cache hit std→RECAL
Llama-3.2-1B (α=.48, a correct sample for 62% of questions)					
1,000	.1	.959	.515 → .589	4.5 → 9.4	.61 → 1.00
1,000	.05	.980	.515 → .601	4.5 → 10.5	.61 → 1.00
2,000	.1	.959	.512 → .585	4.4 → 9.0	.66 → 1.00
2,000	.05	.980	.512 → .603	4.4 → 10.7	.66 → .96
2,419	.1	.948	.511 → .577	4.4 → 8.4	.61 → 1.00
2,419	.05	.974	.511 → .603	4.4 → 10.7	.61 → 1.00
Llama-3.2-3B (α=.41, a correct sample for 69% of questions)					
1,000	.1	.958	.589 → .655	4.0 → 7.5	.71 → .89
1,000	.05	.979	.589 → .663	4.0 → 8.3	.71 → .98
2,000	.1	.967	.586 → .639	3.8 → 6.4	.81 → .95
2,000	.05	.984	.586 → .670	3.8 → 8.5	.81 → 1.00
2,419	.1	.973	.586 → .652	3.8 → 7.0	.83 → 1.00
2,419	.05	.986	.586 → .664	3.8 → 8.1	.83 → 1.00

at n=2,419 is 0.948, so 5% of offsets deploy different rules. The atoms in this score (repeated answers, the no-correct-answer value) are handled by deterministic jitter.

F.5 SELECTION, SEEDS, AND SCORE-UNIT CHECKS

Observable selection. On the synthetic image pool, among M recalibrations, each on n_{cal}=5,000 points drawn from analyst A's half, an adversary deploys the one with the lowest coverage on a 2,000-point validation set from the other half. Coverage is then scored on the held-out 10,000-point test split, with 200 trials per setting and $\rho \in \{.1, .01\}$, $M \in \{20, 50\}$ (Table S18). Validation coverage is nondecreasing in the threshold, so the rule deploys a smallest candidate threshold, up to ties in validation coverage (broken by index). Three comparators isolate mechanisms. *Standard conformal with its rank raised* so that its honest test coverage approximately matches RECAL's is an oracle diagnostic that isolates the effect of excess coverage, since its rank is matched on the test split. A *training-conditional rank* at $\delta = 1/M$ (Vovk, 2012; Bian & Barber, 2023) is the natural defense when only post-selection coverage matters. At this δ a union bound over the M candidates is vacuous, so it carries no guarantee after selection, and the table reports what it achieves. RECAL$^+$ here denotes that training-conditional rank rounded on the shared grid, the exact-rank variant of the rule of Proposition 3, which uses a DKW margin instead. At ρ=.1 the selection drop of RECAL (.0070) is close to that of the coverage-matched standard rule (.0059), so the excess coverage explains the protection. At ρ=.01, where the M candidates deploy 1.02 distinct maps on average, RECAL loses .0005 against .0029 for the coverage-matched rule at M=20 (.0016 against .0036 at M=50): once identity is near one there is nothing left to select, since a ρ-replicable rule run with one shared seed has all M candidates equal except with probability at most $(M-1)\rho$, by a union bound over the pairs $(1, j)$. Selection among conformal sets is studied by Hegazy et al. (2026), reuse of one calibration set across many predictions by Balinsky & Balinsky (2025), and selective inference more broadly by Berk et al. (2013); Fithian et al. (2014); Taylor & Tibshirani (2015).

Selection with the population-coverage rule (in-pool protocol). With n=5,000, M=20 redraws, 300 trials and a pre-specified per-trial offset, the adversary deploys the draw with minimal population coverage. At ρ=0.1, standard calibration selects .8920 (standard deviation .0024 across trials) against an honest .9003, and RECAL selects .9170 against an honest .9239. Two or more distinct RECAL outputs appear in 30.7% of trials. The union bound $(M-1)\rho$=1.9 exceeds one, so it guarantees nothing at ρ=0.1, and the selected coverage stays above .90 because of the excess coverage from rounding. At ρ=0.01, two or more distinct outputs appear in 1.3% of trials against the bound 19%, and RECAL selects .9793 against an honest .9797.

Seed-conditional replicability (in-pool protocol). On the synthetic image pool at n=25,000 and ρ=.1, we used 200 fixed offsets with 200 pairs each. The conditional mismatch $m(u)$ has mean

Table S18: Observable selection under the master protocol on the synthetic image pool (M candidate recalibrations, each on n_{cal}=5,000 points from analyst A's half, a 2,000-point validation set from the other half, test coverage on the held-out 10,000-point split, 200 trials). Coverage honest→selected, mean size honest→selected on 40 trials, and the mean number of distinct deployed maps among the M candidates. Standard errors of the coverage entries are $\leq .002$. The raised-rank row is an oracle diagnostic (rank matched on the test split).

Method	Coverage	Size	Distinct maps
ρ=.1, M=20			
Standard split conformal	.899 → .891	15 → 14	19.3
Standard, rank raised to RECAL's coverage	.922 → .916	21 → 19	19.2
Training-conditional rank, δ=1/M	.906 → .900	16 → 15	19.3
RECAL	.922 → .915	21 → 19	1.3
RECAL$^+$	.928 → .923	24 → 21	1.3
ρ=.1, M=50			
Standard split conformal	.899 → .889	15 → 13	45.4
Standard, rank raised to RECAL's coverage	.922 → .915	21 → 18	44.8
Training-conditional rank, δ=1/M	.907 → .900	17 → 15	45.2
RECAL	.922 → .914	22 → 18	1.3
RECAL$^+$	.929 → .923	25 → 22	1.4
ρ=.01, M=20			
Standard split conformal	.899 → .891	15 → 14	19.3
Standard, rank raised to RECAL's coverage	.976 → .973	68 → 60	18.5
Training-conditional rank, δ=1/M	.906 → .900	16 → 15	19.3
RECAL	.978 → .978	137 → 137	1.0
RECAL$^+$	.980 → .979	137 → 137	1.0
ρ=.01, M=50			
Standard split conformal	.899 → .889	15 → 13	45.4
Standard, rank raised to RECAL's coverage	.976 → .972	68 → 58	41.1
Training-conditional rank, δ=1/M	.907 → .900	17 → 15	45.2
RECAL	.979 → .977	137 → 137	1.0
RECAL$^+$	.980 → .979	137 → 137	1.0

0.0685, median 0, 90th percentile 0.33 and maximum 0.555. It exceeds ρ for 19.5% of seeds and .5 for 1.5%. The two-cell Gaussian form of Proposition 4(ii), $2\Phi(d/\sigma_{S,n})\{1-\Phi(d/\sigma_{S,n})\}$, with d the distance from the pool quantile to the nearest grid boundary ($\sigma_{S,n}$=0.0022, β=0.0320), predicts the per-seed values with R^2=0.98. Within the Gaussian approximation, truncating to the two cells nearest the boundary is accurate to 10^{-12} at this width. This does not bound the error of the normal approximation itself. The Markov bound $\mathbb{P}_u[m > t] \leq \pi_2/t$ holds at every t. For an audit that fixes one public seed, the ex-ante ρ is an average of a bimodal quantity. The centred grid of Theorem 3 replaces this seed dependence with a dependence on the pilot's location (Table S6).

Sample-complexity frontier. We use $\varepsilon \in \{.01, .02, .04, .08\}$, with the width $\beta = \varepsilon/(2\hat{f})$ fixed by ε alone, n on a 17-point geometric grid from 250 to 64,000, and 600 pairs per (ε, n) setting. For each ε, the logarithm of the measured mismatch is regressed on $\log n$ over the unsaturated window in which the mismatch lies in $[0.02, 0.45]$. The measured decay slopes are $-0.49, -0.50, -0.43, -0.49$, against the theoretical $-\frac{1}{2}$. The crossing $n^*(\varepsilon, \rho)$ solves the fitted line at mismatch ρ, and so is an empirical *mismatch* frontier. With $\beta = \varepsilon/(2\hat{f})$, the heuristic mean rounding inflation is $\varepsilon/4$ when the density is locally constant and $\hat{f} \approx f(q)$. The measured mean inflation is .0027/.0053/.0098/.0175 for ε = .01/.02/.04/.08, against $\varepsilon/4$ = .0025/.005/.01/.02. The width alone does not certify a high-probability coverage cap. Six of the twelve (ε, ρ) crossings fall inside the measured grid and enter the frontier fit, with slope 2.09 in $1/(\varepsilon\rho)$ and a CI of $[1.99, 2.28]$ from resampling the pairs within each setting and refitting. Two fall just beyond 64,000 (1.10 and 1.28×10^5). They are shown as open markers and excluded. Four fitted crossings fall beyond twice the largest measured n and are not reported. The fitted exponent describes this algorithm's mismatch frontier on this testbed. It is consistent with the $(\varepsilon\rho)^{-2}$ rate of Corollary 3 and Theorem 4 but is not an empirical proof of the lower bound. The $\alpha(1-\alpha)$ factor is probed separately below, and the κ^2 sample-size factor is not probed. On the real ImageNet logits ($\varepsilon \in \{.02, .04, .08\}$, $\rho \in \{.05, .1, .2\}$, 400 pairs per setting), the

decay exponents are $-0.61/-0.56/-0.54$ and the six resolved crossings give a fitted slope of 1.83 (Figure S6).

Table S19: **Finite-sample effect of the density ratio.** Coverage inflation of RECAL at the plug-in width, divided by its value $f(q)\beta_n/2$ when the density is flat, on the designed law of the paragraph below (α=.1, ρ=.1, flat window of half-width ζ=.03, 300 pairs per setting and 120 at the two largest n, mismatch integrated over the offset and inflation averaged over eight fixed offsets). A ratio of 1 means that the density ratio κ has no effect. A dot marks a setting that was not run. Last row: the width relative to the flat window, β_n/ζ, at κ=2.

κ	n=2,000	8,000	32,000	128,000	512,000	2,048,000
1	0.99	0.99	0.99	1.00	·	·
2	1.22	1.40	1.55	1.30	1.03	0.98
4	1.57	2.00	2.55	2.30	1.22	1.01
8	1.96	2.81	4.03	4.70	·	·
β_n/ζ	18.02	9.01	4.51	2.25	1.13	0.56

Finite-n κ test. A designed score law at α=.1 has unnormalized density 1 on a window of half-width ζ=.03 around the target quantile, κ on the rest of the margin window of half-width .1, and 1 elsewhere (a long left segment and a right segment of length .4), normalized so that 10% of the mass lies above the quantile. The normalized density at the quantile is $f(q) = 1/(.7\kappa + 4.3)$, which is .200/.175/.141/.101 for $\kappa = 1/2/4/8$, so β_n/ζ crosses one only near $n = 6 \times 10^5$. Scores are fresh continuous draws. The plug-in uses $f(q)$ ($\hat{\kappa}$=1, plug-in width). Table S19 reports inflation relative to the κ-free value $f(q)\beta_n/2$. The ratio rises with κ while the grid cell exceeds the flat window and returns toward one as the cell shrinks below it (0.98× at $\beta_n/\zeta = 0.56$ for κ=2, 1.01× at 0.70 for κ=4). This supports a local-flatness explanation: once a grid cell fits inside the flat window, the coverage inflation of the plug-in width no longer depends on κ. The grid reaches $\beta_n = .56\zeta$, just above $\zeta/2$, the width below which a cell that starts within $\zeta/2$ of q stays inside the flat window. Mismatch is at most .091 in every setting. Where $\beta_n/\zeta \le 8$ it lies within .012 of $\rho\sqrt{2/\pi} \approx .080$, the locally flat limit of Proposition 2 at the plug-in width, since the width is set by the density at the quantile. At small n and large κ it is lower (down to .037), because the raw cutoff's spread then reaches the denser shoulders of the law.

α-dependence probe (in-pool protocol). At a grid width β^* fixed at its value for α=.1, $\beta^* = 2B_n^+(.1)/(\hat{f}\rho)$ with ρ=.1 ($\hat{f}$=1 on the rank-transformed pool), and n=25,000 on the synthetic image pool, mismatch was measured over 800 pairs at each $\alpha \in \{.05, .1, .2, .3, .4, .5\}$ in two designs. On the raw APS score, whose pilot density at the working quantile varies 22× over the sweep ($\hat{f}(q_\alpha)$ from 0.45 to 9.7), observed mismatch falls from .069 to .0025. It matches the plug-in prediction $B_n^+(\alpha)/(\hat{f}(q_\alpha)\beta^*)$ with through-origin proportionality 0.97 and $R^2 = .99$. On a rank-transformed copy of the pool ($f \equiv 1$ up to ties), which isolates the $\alpha(1-\alpha)$ factor, mismatch rises overall from .028 to .073 (.043/.058/.050/.044 at the intermediate levels, so the rise is not monotone). It lies below the theoretical bound $B_n^+(\alpha)/\beta^*$ at every level (.037 to .083). A fit against $\sqrt{\alpha(1-\alpha)}$ gives $R^2 = .54$ at 800 pairs, with binomial standard errors $\approx .008$.

Asymptotic-limit check. This check uses the width $2B_n^+/(\hat{f}\rho)$, i.e. $c_0 = 2$ in Proposition 2. On the rank-transformed pool ($f \equiv 1$, so the pilot ratio drops out) at n=1.6 × 10^5, ρ=.1 and 10^4 pairs, mismatch/$\rho = .382 \pm .019$ against the limit $\sqrt{2/\pi}/2 \approx .399$. Inflation×$\rho/B_n^+$ equals 1.001–1.005 from n=10^4 on (limit 1), and the width with $c_0 = \sqrt{2/\pi}$ achieves mismatch .100 at the ρ=.1 target. On the raw APS score the pilot ratio enters as predicted: inflation×ρ/B_n^+ rises from .56 to .93 toward $f(q)/\hat{f} = 1.04$, and mismatch/ρ moves toward .399 $\hat{f}/f(q) = .38$ as $\beta_n \downarrow 0$. The κ test uses fresh continuous draws, whereas the α-probe and this check use rank-transformed copies of the finite pool.

Analyst-side location shift (synthetic image pool, in-pool, plug-in width β=.0320, 200 pairs per shift). Shifting analyst B's scores by $\theta \in \{0, .25, .5, 1, 2\} \times \beta$ gives identity .915/.745/.54/.045/0, against the bound $1 - |\theta|/\beta - \pi_2$ of Proposition 4(iv) with the asymptotic $\pi_2 \approx .8\rho$, .92/.67/.42/0/0 (at $\theta = 0$ the measured value is within one binomial standard error, .020, of the bound), while B's

coverage of its own population stays at .912–.913. Identity is lost once the two quantiles differ by a cell, but validity is not.

Unequal calibration sizes and score mismatch (synthetic image pool, ρ=0.1, in-pool). With β set by the smaller analyst's rule, $(n_A, n_B) = (25\text{k}, 5\text{k})$ measures mismatch $.065 \pm .012$ against the leading term .102 of Corollary 4, and $(25\text{k}, 1\text{k})$ measures $.053 \pm .011$ against .085 (400 pairs each, both analysts' coverage $\geq .92$). When analyst B's scores are perturbed within sup norm η_s, the worst case (a constant $+\eta_s$ shift) measures $.087/.123/.190$ at $\eta_s/\beta = .05/.1/.2$. The additive bound of Corollary 5 is base$+\eta_s/\beta = .143/.193/.293$ (300 pairs, base $.093 \pm .017$). Mean-zero noise of the same sup norm leaves mismatch at .080–.110. The population-mismatch term and the membership bound of Corollary 5 are not probed.

Score factorial on real ImageNet. Table S20 runs standard and RECAL calibration across LAC ($1 - p_y$), APS, RAPS (k_{reg}=5, λ_{reg}=.05) and SAPS, the sorted adaptive prediction sets of Huang et al. (2023), which keep the largest softmax probability and add a weight of .2 per rank below it, on the same logits (in-pool protocol, n=25,000, ρ=.1, 100 pairs). The premium is driven by the expected label mass of the part of a cell that rounding adds above each score's quantile (about half a cell's mass when the label count is locally linear). RAPS has the smallest premium in this comparison and SAPS the largest.

Table S20: Score factorial on real ImageNet logits (in-pool protocol, n=25,000, ρ=.1, plug-in width, 100 pairs, sizes on the first 25 pairs over 10,000 evaluation points). LAC: one minus the softmax probability of the label. SAPS: sorted adaptive prediction sets (Huang et al., 2023).

	Standard			RECAL ρ=.1			
Score	Ident.	Coverage	Size	Ident.	Coverage	Size	Premium
LAC	.00	.8996	3.05	.87	.9084	3.48	+14%
APS	.00	.9001	34.60	.96	.9138	41.11	+19%
RAPS	.00	.8999	5.55	.91	.9126	5.94	+7%
SAPS	.00	.8996	3.36	.96	.9131	4.46	+33%

Table S21: In-pool protocol, fixed density window (60 pairs at n=25,000 for the image pool and n=800 for digits, and 40 pairs at n=5,000 resampled from a 5,000-context pool for Llama). **The replication spectrum** (Definition 1) at ρ=0.1. "Churn" is the mean $|\mathcal{C}(X) \triangle \mathcal{C}'(X)|$ per test point, marginal and conditional on the *pair* mismatching (with the number of mismatched pairs). "bag-10" averages ten bootstrap-resample conformal quantiles of each analyst's own sample, and "derand." is standard split conformal with the shared jitter. Identities of about .02–.05 for the baselines are ties from resampling a small finite pool with replacement.

	Method	Ident.	Agree	Churn (marg.)	Churn | pair mism. (#)
synthetic image	derand.	.02	.66	0.36	0.37 (59)
	bag-10	.00	.63	0.39	0.39 (60)
	RECAL .1	.97	.97	0.19	5.82 (2)
digits (n=800)	derand.	.05	.98	0.019	0.02 (57)
	bag-10	.00	.98	0.017	0.02 (60)
	RECAL .1	.93	.93	0.60	9.01 (4)
Llama-3.2-1B	derand.	.00	.51	5.0	5.0 (40)
	bag-10	.00	.52	4.8	4.8 (40)
	RECAL .1	.90	.92	15.1	151 (4)

The replication spectrum. Table S21 shows how identity affects other criteria. Averaging bootstrap quantiles leaves agreement and churn where they were and identity at 0. RECAL concentrates disagreement into rare events of whole grid cells, so churn conditional on a mismatched pair rises (image $0.37 \to 5.82$, digits $0.02 \to 9.01$, Llama $5.0 \to 151$ labels) while mismatches become rare. On the small digits label space, where standard calibration already agrees at .98 of inputs, marginal churn rises from .019 to .60. Identity is a different target from pointwise agreement, and it can increase churn where the baseline disagreement is smaller than one cell. Each row uses the same pairs, offsets and test inputs for all columns, so identity $\leq$ agreement holds row by row.

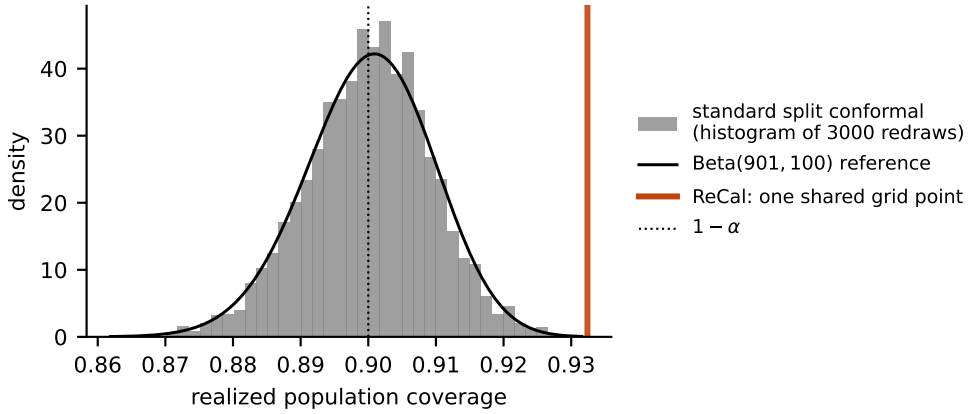


Figure S2: Realized population coverage across calibration redraws (synthetic image pool, in-pool protocol, n=1,000, ρ=.1). The histogram of 3,000 standard calibrations is compared with the continuous-law $\mathrm{Beta}(901, 100)$ reference (the finite pool has atoms, so its law is discrete), and RECAL concentrates the coverage on one shared grid point in this run, above $1 - \alpha$ (dotted). The guarantee is marginal validity: the realized atom can lie below $1 - \alpha$.

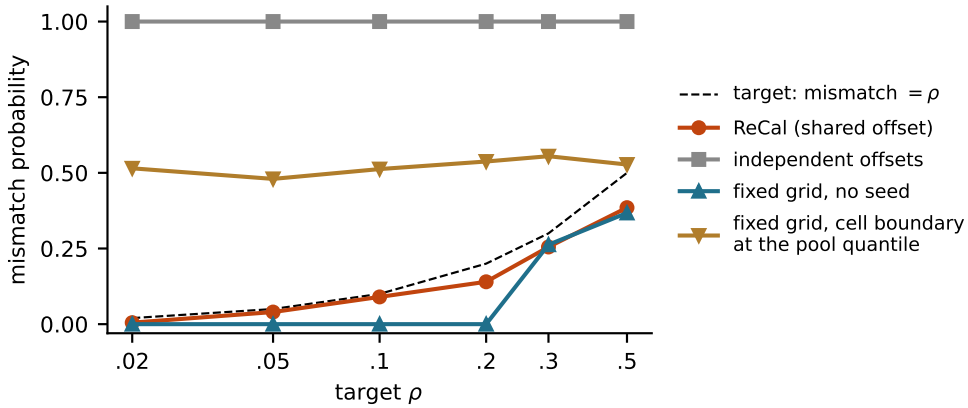


Figure S3: Mismatch probability of the deployed thresholds against the target ρ (synthetic image pool, in-pool protocol, n=25,000, plug-in width, from .160 at ρ=.02 to .0064 at ρ=.5, 400 pairs per ρ). RECAL stays at or below the dashed target line. Independent, unshared offsets give mismatch 1. A fixed grid without a seed has mismatch 0 for $\rho \leq .2$, because this pool's quantile lies well inside a cell, and .26 and .37 at ρ=.3 and .5, where the cells are narrower. Moving the fixed grid so that a cell boundary sits at the pool's quantile drives its mismatch to about $1/2$ at every ρ (Lemma 2 and Proposition 4(ii)).

F.6 FIGURES, COMPUTE AND REPRODUCIBILITY

In-pool figures. Figure S2 shows how upward rounding collapses the Beta law of realized coverage onto a grid atom, Figure S3 shows the achieved mismatch against the target ρ, including the failure of unshared offsets and of a fixed grid with a cell boundary at the pool's quantile, and Figure S5 shows set size and coverage against ρ. Figures S4 and S6 are the frontiers discussed above.

Monte Carlo precision. Offset-integrated identities are averaged as bounded real-valued observations. Sampled-offset streams are Bernoulli, and an estimate $\widehat{p}$ from m independent pairs has standard error $\sqrt{\widehat{p}(1-\widehat{p})/m}$, for example .022 at $\widehat{p} = .95$, $m = 100$. On digits at n=800 and ρ=.1, $\hat{\kappa}$=1, the 600-pair table stream gives mismatch $.113 \pm .013$ and the independent 400-pair safety stream gives $.075 \pm .013$. The two are 2.1 combined standard errors apart and bracket the target, reflecting plug-in variability at small n. $\hat{\kappa}$=1.5 and 2 give $.050 \pm .011$ and $.048 \pm .011$ on the safety stream. A zero-failure outcome over 100 pairs certifies only a rate $\leq .03$ at 95% confidence.

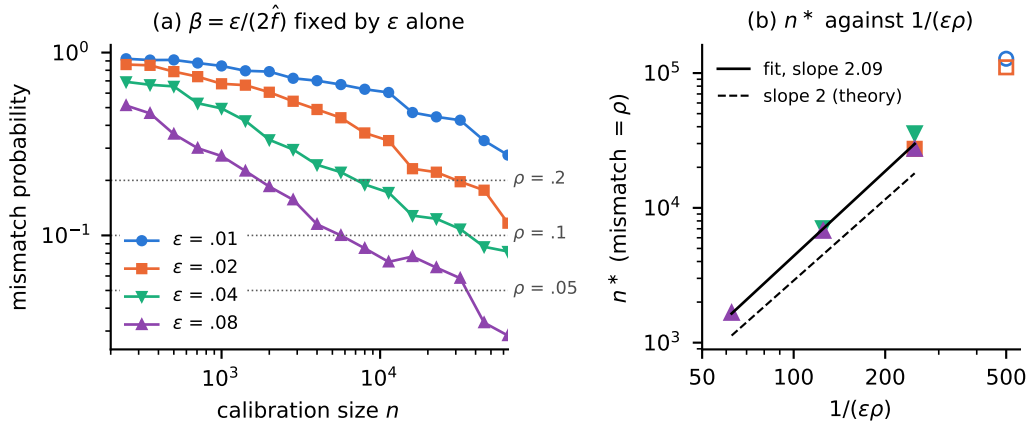


Figure S4: **Measured mismatch frontier** (synthetic image pool, in-pool protocol), with $\beta = \varepsilon/(2\hat{f})$ fixed by ε alone. (a) Mismatch against n, with decay exponents $-.49, -.50, -.43, -.49$ across ε (predicted $-\frac{1}{2}$). Small-ε curves saturate while β is small relative to the threshold spread. (b) n^* against $1/(\varepsilon\rho)$, with markers as in (a), fitted slope 2.09 (CI $[1.99, 2.28]$), consistent with the rate of Corollary 3 and Theorem 4. Crossings with equal $\varepsilon\rho$ share an abscissa: three of the six fitted crossings lie at $1/(\varepsilon\rho) = 250$ and two at 125. Crossings beyond the measured grid are open markers and are excluded from the fit.

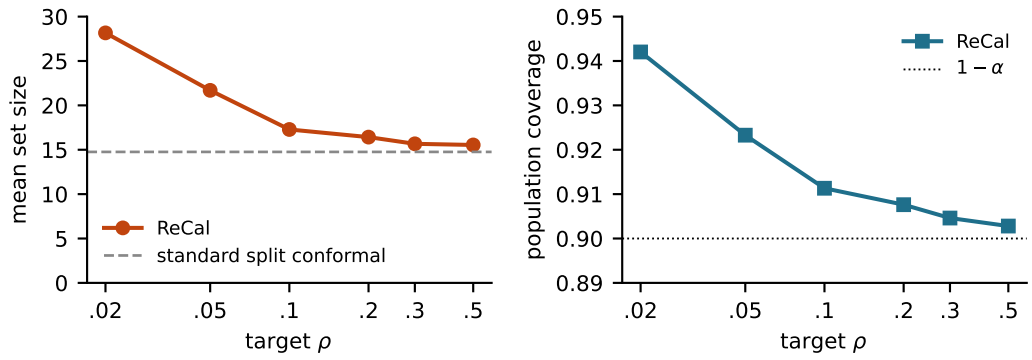


Figure S5: Mean set size (left) and population coverage (right) of RECAL against the target ρ (synthetic image pool, in-pool protocol, n=25,000, plug-in width, 40 calibration draws per ρ and sizes on the first 15). Dashed: mean size of standard split conformal (14.76). Dotted: the nominal coverage $1 - \alpha$. Standard split conformal covers .8995.

Compute. Table S22 lists wall-clock times from the run logs.

Reproducibility map. `v51_master.py` (parts `master`, `selection`, `kappa`, `centered`, `diag`) produces Tables S1 (with the ImageNet rows in its caption), S9, S18 and S19, and the paragraphs on the pilot-centred grid and the cell-mass diagnostic, in `v51_results.json`. A rerun of its master part in a fresh directory reproduced the result file exactly. A rerun of `v8_extras.py --part rho` from a clean copy of the archive, after the two fetch scripts, reproduced the keys `rho_real_aps`, `rho_real_raps`, `rho_synthetic_aps` and `rho_cifar100_aps` of `v8_results.json` value for value (1,008 values each). `fetch_uci.sh` downloads the UCI Heart Disease and Abalone files that `heart_sites.py` and `v5_experiments.py` read. Both scripts take the exact offset averages from `exact_offsets.py`. `v6_regimes.py` (parts `rho`, `shift`, `pilotsize`, each with the pilot-floor rule) produces Table S5, the positive-gap rows of Table S6, and the pilot-floor and conservative pilot-only rows of Tables S1 and S9 in `v6_results.json`. `v8_extras.py` (parts `rho`, `fresh`, `shift`, and `--pyx` for a pool given as released probabilities) produces Tables 2, S2, S3, S4 and S8, the negative-gap rows of Table S6, the RAPS columns of Table 2, and the deterministic-grid and search baselines, in `v8_results.json`. Its part `rho` reuses the pilots and pairs

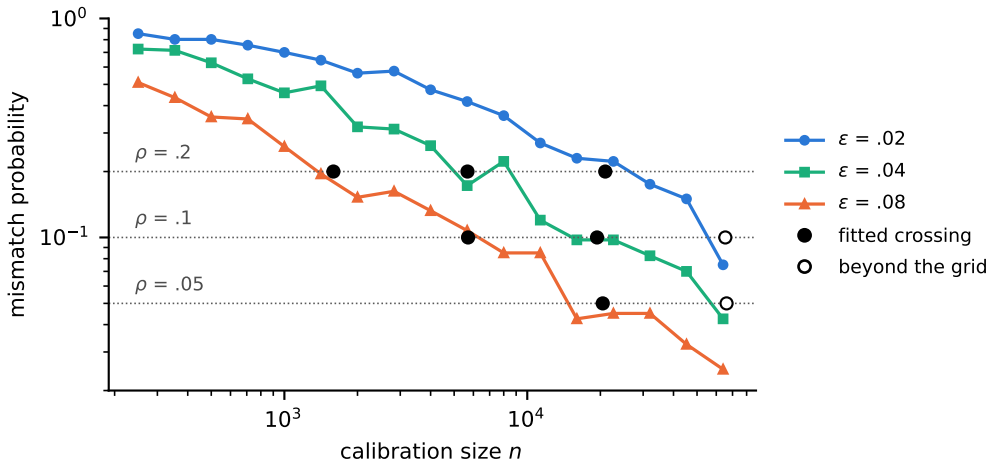


Figure S6: Real-ImageNet mismatch against n at fixed $\beta = \varepsilon/(2\hat{f})$ (in-pool protocol, 400 pairs per point). Dotted lines mark ρ. Black dots are the crossings $n^*(\varepsilon, \rho)$ of the fitted lines that enter the frontier fit (slope 1.83, CI $[1.79, 1.87]$), and open circles are fitted-line crossings just beyond the measured grid, excluded from the fit.

of `v6_regimes.py` and reproduces its identities and coverages at $\rho \in \{.1, .05, .01\}$ to all digits, with sizes now measured on all 1,000 pairs. `fresh_draw_check.py` produces Table S7 in `fresh_draw_check.json`, `v8_fresh_large_rho.py` the fresh draws at $\rho \geq .2$ quoted above, and `certificates.py` the certified widths, ranks and exact identities quoted throughout. The part `rho` of `v6_regimes.py` reuses the master protocol's pilots and pairs, and at ρ=.1 its score-unit identities and coverages equal `v51_master.py`'s to all digits. `design_check.py` gives the two-design comparison above, `finite_pool_model.py` the Gaussian finite-pool model of the floor's identity, `shift_diagnostics.py` the density ratios and pilot densities quoted for the shifted pilot, and `make_fig_v8.py` draws Figures 1, 2 and 3 from `v6_results.json` and `v8_results.json`. `make_figures_nodigits.py` draws Figures S2 to S5 from `results.json`, `scale_curve.py` draws Figure S1 from the Pythia sweep files, and `make_fig7.py` draws Figure S6 from `v5_results.json`. The rows of Table S10 come from an earlier version of `v51_master.py`, run with `--logits` (key `master_real` of `v51_results_cifar100_*.json`). That version predates the coverage-unit rules and the exact offset averages and is not in the archive, so the shipped script, run on the same logits, is not guaranteed to reproduce these rows digit for digit. The `master_synthetic` key of these files is a copy of an earlier synthetic-pool run and is not used in the paper. `b5_llm_raps.py` produces Tables 3, S11 and S12 (the integrated cell-mass predictions are in the files `b5_*_diag.json`), `cpu_folktables_sites.py` and `cpu_folktables_regimes.py` produce Tables S14 to S16, `cpu_conformal_lm_qa.py` produces Table S17, and `experiments.py`, `v5_experiments.py`, `heart_sites.py`, `alpha_sweep.py`, `adaptive_check.py` and `scale_curve.py` produce the in-pool results.

Table S22: Compute (one workstation with two NVIDIA RTX 4090 GPUs of 24 GB and 20 CPU cores). All calibration experiments run on one CPU core. Wall-clock times from the run logs.

Experiment	Script	Hardware	Time
Master protocol, synthetic pool (five parts)	`v51_master.py`	CPU	8 min
Master protocol, real ImageNet	`v51_master.py --pool real`	CPU	3 min
Certified widths, ρ-sweep, pilot size, shifted pilot, design check	`v6_regimes.py`, `certificates.py`, `design_check.py`	CPU	4 min per pool
Seven-level ρ-sweep with baselines (per pool and score), reverse shift, fresh-draw grid, CIFAR-100 sweep (per score)	`v8_extras.py`	CPU	8 / 1 / 11 / 1 min
Public ResNet-50 ImageNet logits, CIFAR-100 probabilities and UCI files	`fetch_imagenet_pyx.sh`, `fetch_cifar100_pyx.sh`, `fetch_uci.sh`	network	3 min
Fresh-draw checks, shift diagnostics, finite-pool model	`fresh_draw_check.py`, `v8_fresh_large_rho.py`, `shift_diagnostics.py`, `finite_pool_model.py`	CPU	3 min
Figures 1 to 3	`make_fig_v8.py`	CPU	1 min
CIFAR-100 fine-tune, ResNet-18 (3 epochs)	`gpu_cifar100_two_archs.py`	GPU	47 s
CIFAR-100 fine-tune, ViT-B/16 (2 epochs)	`gpu_cifar100_two_archs.py`	GPU	4 min
Master protocol on each CIFAR-100 logit file (earlier version of the script)	`v51_master.py --pool real --logits FILE`	CPU	2 min
Language-model logit extraction, 10k–40k contexts	`extract_logits.py`	GPU	5–25 min per model
Language models, master protocol, per model	`b5_llm_raps.py`	CPU	8–21 min
Answer sampling, Llama-3.2-1B / 3B, K=20	`gpu_lm_sample_answers.py`	GPU	8 / 21 min
Answer-set cache scoring	`cpu_conformal_lm_qa.py`	CPU	4 min
Census sites and shifted-pilot census regime	`cpu_folktables_sites.py`, `cpu_folktables_regimes.py`	CPU	10 / 1 min
In-pool experiments and frontier	`experiments.py`, `scale_curve.py`, `v5_experiments.py`	CPU	11 min
Clinical sites and α-probe	`heart_sites.py`, `alpha_sweep.py`	CPU	2 min
Asymptotic-limit check	`adaptive_check.py`	CPU	4 min
Total			about 3 GPU-hours and 3 to 4 CPU-hours