The Iterated-Logarithm Price of Post-Hoc Threshold Selection

Anonymous authors
Paper under double-blind review

ABSTRACT

The standard way to deploy a selective predictor is to calibrate it, look at the risk coverage curve, pick the operating point that looks best, and report the risk certificate at that point. Every widely used certificate such as Clopper–Pearson, Hoeffding, empirical Bernstein, betting bounds, and Learn-then-Test is valid only at a threshold fixed *before* the calibration data are seen, so none carries a uniform guarantee under it. Some happen to survive specific attacks at specific scales anyway, but by accident rather than by design. We give a certificate built for the workflow directly. Sorting items by a confidence score is a function of the covariates alone, so it leaves the losses conditionally independent; the resulting *sorted filtration* turns the risk coverage curve into a martingale problem, and anytime valid concentration applies directly. This yields a band U_k that covers the realized risk of every prefix simultaneously, for an arbitrary score, with no assumption that the score is calibrated or even informative. And it does so, we show, with the certificate paying no extra price for dependence within blocks specified in advance. The band's $\sqrt{\ln \ln k / k}$ width is not slack: any band valid at every prefix simultaneously must be at least this wide infinitely often, and we check the rate empirically across calibration sizes spanning a factor of 25. Across three language models, seven score families, five datasets and a vision model, the band was violated 0 times in $32{,}300$ replications with exactly known conditional risk and 0 times in 320 genuine decoding replicates. In a deployment protocol that chooses the threshold after seeing the calibration curve, Clopper–Pearson and a betting certificate are exceeded on held-out data 12.7% and 11.6% of the time while ours is exceeded 0 times in $24{,}000$. We trace the betting certificate's failure to an edge case in our own implementation a 32% silent-margin defect under heterogeneous risk against 1% on i.i.d. data and repair it with a one-line, provably valid patch that resolves the failure under a fixed threshold and substantially, though not completely, under post-hoc selection. A feasibility condition states exactly when a population-level version of the same certificate can reach a target risk level, and we report the regime, small calibration sets and mid-accuracy models, where it cannot, along with how much a lower-error subset closes the gap. Code is available at `https://anonymous.4open.science/r/iterlog-risk-E131/`.

1 INTRODUCTION

Selective prediction is deployed by choosing a threshold. A model emits a confidence score, items scoring above a cutoff are answered and the rest are abstained on or escalated, and the operator wants a guarantee on the error rate among the answered items. That guarantee is what makes the system usable in triage, retrieval filtering, or anywhere a wrong answer costs more than no answer at all (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017).

The trouble is in how the threshold actually gets picked. An analyst holds out a calibration set, plots empirical risk against coverage, and reads off whichever cutoff looks best, so the cutoff ends up a function of the calibration data. The certificates were never built for that. Exact binomial intervals (Clopper & Pearson, 1934), Hoeffding (Hoeffding, 1963), empirical Bernstein (Maurer & Pontil, 2009), betting bounds (Waudby-Smith & Ramdas, 2024), Learn-then-Test over a pre-registered grid (Angelopoulos et al., 2025): each is a statement about a threshold fixed *before* the data are seen. Applying one at a data-chosen cutoff is a garden-of-forking-paths error (Gelman & Loken, 2013), and the post-selection inference literature documents at length how badly that error bites (Berk et al., 2013; Taylor & Tibshirani, 2015). None of this reflects carelessness on the practitioner's part. No certificate on offer permits what they're doing.

What we do. We build one that does from an observation that's simple once stated: the permutation that sorts items by score depends only on the covariates and a set of independent tie-breakers, so

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.

conditioning on it leaves the losses conditionally independent. Read the sorted sequence prefix by prefix and you get a filtration along which the centred loss sum is a martingale with bounded increments; and uniform-in-k control of a martingale is exactly what time-uniform concentration inequalities are built to supply (Ville, 1939; Howard et al., 2020; 2021). That gives a band U_k, valid at every prefix length simultaneously, for a score that can come from output probabilities, hidden activations, self-consistency across samples, or a number the model states out loud, and that need not be well-calibrated or even informative. The same argument extends to blocks chosen by the analyst in advance with arbitrary dependence permitted inside a block at no extra price. The one property the proof leans on, a variance-versus-mean inequality due to Bhatia & Davis (2000), holds for a block average as it holds for a single item. That band costs $\sqrt{\ln \ln k/k}$ more than a fixed-k interval would. Our second result says this price can't be negotiated away. In the homogeneous case, where the score carries no information, the law of the iterated logarithm (Hartman & Wintner, 1941) forces any band valid at every k to be at least $\sqrt{2p(1-p)\ln \ln k/k}$ wide infinitely often what simultaneity costs, not an artifact of this construction, checked across four calibration sizes spanning a factor of 25.

What we found. The falsification suite behind these claims spans three language models, seven score families across output, sequence, hidden-state, sampling, and verbalized confidence, five datasets, and a ResNet-18 on CIFAR-10 run through the full CIFAR-10-C sweep. Where possible, ground truth for the conditional risk is exact—read directly off a model's own predictive distribution, not estimated. The headline validity result is a long string of zeros: no violation across 32,300 replications on real activations spanning 33 arms, none across 320 genuine decoding replicates, none across 5,300 synthetic replications that include three separate ablations of the band's internal construction, and none across a sweep of induced within-block correlation up to $\rho = 0.8$ on two unrelated block structures. Three findings surprised us, worth reporting for that reason. When the threshold is chosen after the fact, Clopper–Pearson and a betting certificate blow through their nominal budget on held-out data while Hoeffding and empirical Bernstein don't—the opposite of what we pre-registered, for reasons that are arithmetic, not deep (Section 5.4). Second, chasing the betting certificate's failure led somewhere unexpected: not selection, but a single line in our own implementation that silently returns zero margin under a condition firing thirty times more often on real, heterogeneous-risk data than on i.i.d. synthetic data of the same size. A one-line patch that takes the max with an always-valid fallback closes almost all of the gap, and the part it doesn't close is informative: exactly what the patch's proof doesn't cover. Third, deliberately breaking the theorem's independence assumption broke item-by-item certification right along with it, exactly as it should; certifying at the block level restored validity within the correlation we'd induced, at the same width the block count itself sets, on both a language and a vision benchmark. None of this is free: on the calibration sizes and model strengths studied, the certificate is provably valid throughout but reaches a usable operating point only on the easiest subset we tried, not the pool as a whole (Section 5.2); validity and practical usability are different questions, and we report both.

Contributions.

1. A uniform risk band over the sorted filtration (Theorem 1), valid simultaneously at every coverage level for an arbitrary score, together with a population form that transfers to fresh data (Theorem 3), an explicit feasibility condition (Corollary 1), and a joint risk-and-coverage form (Corollary 2) that comes free from the same argument.
2. A block-robust extension (Theorem 4): the band survives arbitrary within-block dependence at no extra price, once the analyst fixes a block structure in advance, verified by inducing correlation directly and watching item-level certification fail while block-level does not.
3. A matching lower bound (Theorem 2): in the homogeneous regime, any uniformly valid band pays the same iterated-logarithm rate infinitely often, checked empirically across four calibration sizes spanning $25\times$.
4. A three-tier evaluation protocol separating exactly-known conditional risk from estimated risk from held-out measurement, a deployment study built on it, and a complete account of every place we departed from our own pre-registration.

2 SETUP

Let $(X_i, Y_i)_{i=1}^n$ be drawn i.i.d. from a distribution P, and let $\ell_i = \ell(f(X_i), Y_i) \in \{0, 1\}$ be a bounded loss for a fixed predictor f (the martingale arguments behind Theorems 1 and 4 extend directly to $\ell \in [0, 1]$, reading η as a conditional mean; the scalar embedding behind Theorem 3 needs binary $\ell \in \{0, 1\}$). Write $\eta(x) = \mathbb{E}[\ell \mid X = x]$ for the conditional risk and $\eta_i = \eta(X_i)$. A *score* is any measurable $s : \mathcal{X} \to \mathbb{R}$, larger meaning more confident, and we assume nothing else about it: it need not be calibrated, monotone in η, or informative. Draw i.i.d. tie-breakers $\zeta_i \sim \mathrm{Unif}(0, 1)$, independent of everything else, and let π sort the pairs $(s(X_i), \zeta_i)$ in decreasing lexicographic order. For a prefix length $k \in \{1, \ldots, n\}$, define the *realized* (or *transductive*: bound to this specific sample,

not a future one) and *empirical* risk–coverage curves

$$\bar{R}_k \;=\; \frac{1}{k}\sum_{i=1}^{k}\eta_{\pi(i)}, \qquad \widehat{R}_k \;=\; \frac{1}{k}\sum_{i=1}^{k}\ell_{\pi(i)}, \qquad S_k = k\widehat{R}_k. \tag{1}$$

$\bar{R}_k$ is the quantity a certificate should bound—the average conditional risk among the k items the system would actually answer. It isn't observable; $\widehat{R}_k$ is.

Definition 1 (Uniform validity)**.** A band $\{U_k\}_{k\ge k_{\min}}$, with U_k a measurable function of $(\widehat{R}_1,\dots,\widehat{R}_k)$, is δ-*uniformly valid* if $\Pr\big[\,\bar{R}_k \le U_k \text{ for all } k \ge k_{\min}\big] \;\ge\; 1-\delta$.

This is what makes post-hoc selection legitimate. If the event above holds, then for *any* rule $\hat{k}$ that picks a prefix length using the calibration data—or anything else, for that matter—$\bar{R}_{\hat{k}} \le U_{\hat{k}}$ holds on that same event. The certificate gets chosen before the analyst does; after that, the analyst is free to look at everything.

The sorted filtration. Let $\mathcal{F}_0 = \sigma(X_{1:n}, \zeta_{1:n})$ and $\mathcal{F}_k = \sigma(\mathcal{F}_0, \ell_{\pi(1)}, \dots, \ell_{\pi(k)})$. It's i.i.d. sampling, not the weaker exchangeability conformal prediction usually assumes, that makes π safe to condition on: the joint law factors as $\prod_i p(x_i)p(y_i \mid x_i)$, so given $X_{1:n}$ the Y_i are mutually independent with a law depending only on their own X_i. Bare exchangeability doesn't give you that—a pair sequence can be exchangeable while a shared latent factor leaves losses arbitrarily dependent conditional on the covariates.[1] Since π is $\mathcal{F}_0$-measurable, $\ell_{\pi(k)} \mid \mathcal{F}_{k-1}$ is Bernoulli$(\eta_{\pi(k)})$ with $\eta_{\pi(k)}$ itself $\mathcal{F}_0$-measurable, so

$$N_k \;=\; \sum_{i=1}^{k}\big(\eta_{\pi(i)} - \ell_{\pi(i)}\big) \;=\; k\bar{R}_k - S_k \tag{2}$$

is an $(\mathcal{F}_k)$-martingale with increments in $[-1,1]$ and predictable variance $V_k = \sum_{i\le k}\eta_{\pi(i)}(1-\eta_{\pi(i)})$. That's the entire reduction. Sorting looks like a data-dependent move—but it's data-dependent only through the covariates, which is precisely what the martingale structure can absorb.

3 Uniform bands and their price

3.1 The band

For $\lambda \in (0,1]$ write $a(\lambda) = \frac{\lambda}{2(1-\lambda/3)}$, so $a(\lambda) \le 3/4$ throughout, with equality exactly at $\lambda = 1$. Fix $\delta \in (0,1)$ and a geometric base $b > 1$ (we use $b = 2$), and for $j = 0, 1, 2, \dots$ set

$$\delta_j = \frac{6\delta}{\pi^2(j+1)^2}, \qquad \ell_j = \ln(1/\delta_j), \qquad \lambda_j = \min\Big\{1,\ \sqrt{2\ell_j/b^j}\Big\}, \tag{3}$$

so $\sum_{j\ge 0}\delta_j = \delta$. Choosing $\lambda_j \asymp \sqrt{\ell_j/b^j}$ makes the j-th bound tightest around $k \approx b^j$; minimizing over j stitches the pieces into one band that tracks k as it grows.

Theorem 1 (Uniform risk–coverage band)**.** *With probability at least* $1-\delta$*, simultaneously for every* $k \ge 1$*,*

$$\bar{R}_k \;\le\; U_k \;:=\; \min_{j\ge 0}\ \frac{\widehat{R}_k + \ell_j/(\lambda_j k)}{1 - a(\lambda_j)}. \tag{4}$$

Proof idea. Freedman's supermartingale construction applied to (2) at a fixed λ, combined with Ville's inequality, gives a bound at parameter λ holding for every k at once. Applying this at each λ_j with budget δ_j from (3) and union-bounding over j spends exactly δ in total; the minimum over j then preserves validity at every k. Full derivation in Appendix A. □

Two things are worth flagging. The variance bound $\eta(1-\eta) \le \eta$ lets U_k be computed from $\widehat{R}_k$ alone, needing no separate variance estimate—the same self-normalizing trick behind empirical-Bernstein rates (Audibert et al., 2007). And the union over j costs only $\ln(j+1)^2$, which at $j \approx \log_b k$ is $O(\ln\ln k)$: this is where the iterated logarithm enters, and Theorem 2 says it has no choice but to.

Remark 1 (An explicit closed form, verified but not proven here)**.** Let $\ell(k) = \ln\big(\pi^2(\lceil\log_2 2k\rceil + 1)^2/(6\delta)\big)$, the menu entry (3) tuned to scale k. Whenever $k\widehat{R}_k \ge 2\ell(k)$, evaluating (4) at this single

[1] Concretely: $U \sim \mathrm{Unif}(0,1)$ shared across i, X_i i.i.d. $\mathrm{Unif}(0,1)$, $Y_i = \mathbf{1}\{U < X_i\}$ is exchangeable in (X_i, Y_i) but $\mathrm{Corr}(Y_i, Y_j \mid X_{1:n})$ can approach 1.

entry empirically tracks $U_k^{\mathrm{c}} = \widehat{R}_k + 4\sqrt{\ell(k)\widehat{R}_k/k} + 8\ell(k)/k$ closely enough to trade tightness for a closed form (width 0.118 vs. 0.034 at $n = 20{,}000$, $k = 2{,}000$, $\widehat{R}_k = 0.10$, $\delta = 0.05$), verified numerically in Table 3 and Appendix A; the general algebraic link isn't fully closed, so we present U_k^{c} as a practical approximation, not a proven alternative.

3.2 Necessity of the iterated logarithm

Theorem 2 (Iterated-logarithm lower bound)**.** *Consider the homogeneous regime $\eta \equiv p \in (0,1)$ with $s(X)$ independent of ℓ, so $\bar{R}_k = p$ for every k. Let $\{U_k\}$ be any δ-uniformly valid band. Then on an event of probability at least $1-\delta$,*

$$\limsup_{k\to\infty} \frac{U_k - \widehat{R}_k}{\sqrt{2p(1-p)\ln\ln k/k}} \;\geq\; 1. \tag{5}$$

Proof idea. On the validity event $U_k \geq p$ for all $k \geq k_{\min}$, so the width satisfies $W_k \geq p - \widehat{R}_k$. Because the losses here are i.i.d. Bernoulli(p) and the ordering doesn't touch them, Hartman–Wintner (Hartman & Wintner, 1941) pins down the almost-sure fluctuation of $p - \widehat{R}_k$ at exactly this rate. The two facts combine directly into the claim; full details in Appendix A. □

This is a statement about width at large k, not a claim that every valid band must look like ours. What it says is that the $\ln\ln k$ factor is structural: shave a band uniformly by any constant $c < 1$ and it stops being valid. Section 5.3 measures the price of trying anyway, at four calibration sizes rather than one. The same rate shows up for the same reason in best-arm identification (Jamieson et al., 2014) and confidence sequences (Darling & Robbins, 1967; Robbins, 1970; Howard et al., 2021); what's new here is that the sorted filtration puts risk–coverage curves in that same regime to begin with. It's worth being precise about what this does and doesn't rule out. The bound above is specifically about bands that are uniformly valid over the *sorted prefix* $\bar{R}_k$, in the sense of Definition 1. Section 3.3 gives a second route to a post-hoc-usable certificate that is uniform over *thresholds* instead of prefixes, via a fixed, low-complexity function class rather than a sequentially growing one, and it pays a flat $O(n^{-1/2})$ rather than an iterated-logarithm rate. The two routes aren't in tension—they answer different questions, at different costs—but a reader skimming only the title should know the LIL lower bound is a statement about the prefix band specifically, not about every way of making post-hoc threshold selection legitimate.

3.3 From ranks to thresholds

Deployment fixes a score threshold t, not a rank. Let $\phi(t) = \Pr[s(X) \geq t]$ and $\rho(t) = \Pr[s(X) \geq t,\ \ell = 1]$, so the population selective risk is $R(t) = \rho(t)/\phi(t)$, with empirical counterparts $\hat{\phi}(t) = k/n$ and $\hat{\rho}(t) = S_k/n$ on a calibration set of size n, where $k = |\{i : s(X_i) \geq t\}|$.

Theorem 3 (Population band)**.** *Let $\varepsilon_n(\delta) = \sqrt{\ln(2/\delta)/(2n)}$. With probability at least $1-\delta$, simultaneously for every threshold $t \in \mathbb{R}$ with $\hat{\phi}(t) > \varepsilon_n(\delta)$,*

$$R(t) \;\leq\; U^{\mathrm{pop}}(t) \;:=\; \frac{\hat{\rho}(t) + \varepsilon_n(\delta)}{\hat{\phi}(t) - \varepsilon_n(\delta)}. \tag{6}$$

The half-line sets $\{s \geq t\}$ form a VC class of dimension 1, so a DKW/Massart argument (Dvoretzky et al., 1956; Massart, 1990) handles $\hat{\phi}$ directly; $\hat{\rho}$ needs one extra step—it's a sub-distribution, not a distribution function—and Appendix A gives the embedding that extends the same inequality to it. Unlike Theorem 1, this one transfers to fresh data at a cost of $O(n^{-1/2})$ uniformly over t—the price every deployment certificate pays somewhere, not the iterated-logarithm price of Section 3.2: it's a fixed, finite-complexity class being made uniform, not a sequentially growing prefix.

Corollary 1 (Feasibility)**.** *$U^{\mathrm{pop}}(t) \leq \alpha$ at a threshold with coverage $\hat{\phi}(t) = \gamma$ iff $\widehat{R}_k \leq \alpha - (1+\alpha)\,\varepsilon_n(\delta)/\gamma$. In particular $\widehat{R}_k \leq \alpha - (1+\alpha)\varepsilon_n(\delta)$ is necessary, attained only as $\gamma \to 1$.*

At $\alpha = 0.10$, $\delta = 0.05$, this reads $\widehat{R}_k \leq 0.067$ at $n = 2{,}000$ and $\widehat{R}_k \leq 0.089$ at $n = 20{,}000$ (Appendix B). Since $\widehat{R}_k \geq 0$ always, the same inequality forces a coverage floor $\gamma \geq \gamma_{\min} := (1+\alpha)\varepsilon_n(\delta)/\alpha$—33.4% at $n = 2{,}000$—below which no predictor, however accurate, can be certified at this target. A model whose top-k empirical error can't be pushed below 6.7% cannot be certified at the 10% level from 2,000 calibration points, whatever score it uses. That's a fact about sample size, not about the method, and Section 5.2 shows it actually binding.

Corollary 2 (Joint risk–coverage certificate)**.** *On the same event as Theorem 3, simultaneously for every t, $\phi(t) \geq \hat{\phi}(t) - \varepsilon_n(\delta)$.*

This costs nothing extra—the same two-sided DKW bound already controls $|\hat{\phi}(t) - \phi(t)|$, and Theorem 3 only used one direction—and answers what the risk certificate alone doesn't: how much of the population a certified threshold reaches, a floor Section 5.2 finds non-trivial in every deployment replication.

Corollary 3 (Two-sided operating-point certificate)**.** *On the same event as Theorem 3, simultaneously for every t with $\hat{\phi}(t) > \varepsilon_n(\delta)$, the pair $\big(U^{\mathrm{pop}}(t),\ \hat{\phi}(t) - \varepsilon_n(\delta)\big)$ jointly upper-bounds the population risk and lower-bounds the population coverage at t, at no additional cost in δ: Theorem 3 and Corollary 2 already hold on this one event, reported here as a single certificate rather than two results combined by hand.*

3.4 Robustness to within-block dependence

Theorem 1's proof leans on exactly one property of the loss sequence: $\mathrm{Var}(\ell_{\pi(k)} \mid \mathcal{F}_{k-1}) \leq \eta_{\pi(k)}$. That property doesn't ask for independence *inside* a group—only that the group, treated as one unit, behaves like $[0,1]$-valued variable with a conditional mean. So the same proof certifies at the level of blocks the analyst names in advance, under dependence the item-level argument cannot see.

Theorem 4 (Block-robust band)**.** *Partition the population into blocks $B_1, B_2, \ldots$, fixed in advance and mutually independent conditional on covariates, with arbitrary dependence permitted* within *a block. Let η_j^{blk} and $\ell_j^{\mathrm{blk}} \in [0,1]$ denote the within-block mean conditional risk and mean realized loss. Sorting blocks by mean score and applying Theorem 1 to the block-indexed sequence (ℓ_j^{blk}) gives a band valid at every block-count k, at the identical width* (4)*, with no correction for whatever dependence lives inside a block.*

Proof idea. Bhatia–Davis (Bhatia & Davis, 2000) says any $[0,1]$-valued variable with mean μ has variance at most $\mu(1-\mu)$, regardless of its internal structure—which is exactly the bound Theorem 1's proof needs, applied to a block average, not a single item. Full argument in Appendix A.5. □

Read this precisely: what U_k actually bounds is $\frac{1}{k}\sum_{j\leq k} \eta_j^{\mathrm{blk}}$, the unobserved conditional risk of the selected blocks, and specifically its *unweighted* average across blocks—not the item-weighted risk pooled across all items in those blocks. When blocks are equal-sized the two coincide; when they aren't—as in our LLM pool, whose four blocks range from about 2,600 to 11,000 items—they can differ, and it's the block average that's certified, at the width formula for $k = 4$ blocks, not $k = 24{,}000$ items: "no correction for dependence" removes the penalty correlation would add, not the cost of having four statistical units instead of thousands. An analyst who blocks in advance pays the latter cost, not the former; item-by-item certification pays neither correctly, since that formula assumes an independence it doesn't have. Section 5.5 builds exactly this failure and its fix.

4 Experimental design

Theorem 1 is falsifiable, so the study tries to falsify it, but $\bar{R}_k$ depends on η, ordinarily unknown; we use three tiers of ground truth and tag every result with its tier.

Tier A (exact η). The predictor is *sample one answer at temperature T*, so its conditional risk $\eta_i = 1 - p_T(y_i^\star \mid x_i)$ comes straight out of the model's logits, and losses are drawn Bernoulli(η_i). The activations, scores, and item difficulties are all real; only the loss realization is simulated, drawn from the model's own predictive distribution. Because η is exact here, so is $\bar{R}_k$, and any violation rate above δ falsifies the theorem outright. The same tier covers the vision model too, with a matched-control arm on clean CIFAR-10 (Section 5.1), and reappears under a deliberately *broken* independence assumption in Section 5.5, via a Gaussian-copula resampling that keeps each item's marginal η_i fixed while inducing controllable within-block correlation.

Tier B (estimated η). On TriviaQA (Joshi et al., 2017) we decode 32 genuine samples per question at $T = 1$ and take $\hat{\eta}_i$ to be the fraction that are wrong. Each draw is then checked, in turn, against the leave-one-out $\hat{\eta}$ curve built from the other 31. Nothing here is simulated; the cost is resolution, since 32 draws can't resolve a 5% rate.

Tier C (held-out measurement). The deployment protocol: split into n_{cal} calibration and $\geq 3{,}000$ held-out items, sort the calibration set, pick a prefix length by some selection rule, convert it to a score threshold, apply that threshold to the held-out items, and check whether the held-out error exceeds the certificate. No η needed; the held-out mean is ground truth by construction.

Models, data, scores. Pythia-410M, Qwen2.5-1.5B-Instruct, and Phi-3.5-mini-instruct are scored on a 24,000-item pool combining ARC, OpenBookQA, and CommonsenseQA, plus MMLU as a separate group; options are ranked by length-normalized log-likelihood (Gao et al., 2023) (model, dataset, and pool-accuracy details in Appendix E). We test seven score families: maximum softmax probability (Guo et al., 2017), predictive entropy, sequence log-probability, two hidden-state variants (centroid distance and a k-NN score; Azaria & Mitchell, 2023; Burns et al., 2023), self-consistency agreement (Wang et al., 2023), and verbalized confidence (Tian et al., 2023; Kadavath et al., 2022). The first five run across Tier A and the deployment protocol; verbalized confidence is parsed on $8{,}000$ prompts per instruct model and evaluated in Tier A; self-consistency appears only in Tier B, where genuine repeated decoding makes it meaningful. A ResNet-18 trained on CIFAR-10 supplies a Tier-A matched-control check on maximum softmax probability (Section 5.1), the exchangeability stress test via CIFAR-10-C, and, with the pool's four datasets, one of the two block structures in Section 5.5.

Baselines. Clopper–Pearson at a fixed k (Clopper & Pearson, 1934); Hoeffding (Hoeffding, 1963); Maurer–Pontil empirical Bernstein (Maurer & Pontil, 2009); a betting/WSR upper confidence bound (Waudby-Smith & Ramdas, 2024); Learn-then-Test with Bonferroni-corrected grids and a fixed-sequence variant (Angelopoulos et al., 2025); and conformal risk control, whose guarantee is in expectation rather than high-probability (Angelopoulos et al., 2024). We also build *transfer-corrected* variants of Clopper–Pearson and Hoeffding (Section 5.4), giving each the same population-transfer allowance Theorem 3 pays for—exact concentration at a single threshold, not the calibration prefix—so the deployment comparison is fair.

Pre-registration. Verdict thresholds, $\delta = 0.05$, $k_{\min} = 100$, target level $\alpha = 0.10$, LTT grid sizes, and all seeds were fixed before any arm ran; $1{,}000$ replications give 98.2% power against a true rate of 0.08. Confidence intervals are exact Clopper–Pearson unless stated otherwise, and p-values are one-sided binomial against δ. Section 5.4 reports where the criterion wasn't met, and Appendix C lists every deviation, including a second wave run to stress-test the first wave's conclusions.

5 Results

5.1 Validity

A violation anywhere below would flag a bug in our implementation, not a gap in the proof: Theorem 1 is already established (Section 3), so this checks code against proof, not an open question—exactly the check a skeptical reader should want run at scale, which is why we ran it $32{,}300$ times rather than take it on faith. Theorem 1 never once broke (band shape and width illustrated on one replication in Figure 4, Appendix H). On real activations with exact η, the violation rate is 0.0000 across all 33 Tier-A arms: three models times two dataset groups times five scores, plus verbalized confidence on both instruct models, plus a matched-control arm on clean CIFAR-10 under the identical protocol. Each ran $1{,}000$ replications, or 300 for the CIFAR arm, for $32{,}300$ total and an exact upper confidence limit of 0.0037 per LLM arm (0.0122 for CIFAR, on its smaller count; every arm individually is in Appendix D). On synthetic planted-risk worlds the rate is 0.0000 over $1{,}000$ replications and stays there under several ablations of the band's own construction—score quantization, an alternative geometric base, an alternative δ-allocation across the menu (3)—with resulting band widths in Table 3, Appendix D. The two-sided version at $\delta/2$ per side violates at 0.0020, as it's supposed to.

Tier B adds real decoding noise on top. Across 10 (model, score) arms and 32 genuine decode replicates each—320 total, none simulated—there were zero violations, raw and resolution-adjusted alike. The per-arm interval is $[0.0000, 0.1089]$, which is admittedly wide: 32 draws can't pin down a 5% rate, and we're not pretending otherwise. Tier B corroborates; Tier A is what actually falsifies.

Figure 8 (Appendix H) shows the bands themselves. Hidden-state scores buy a genuinely tighter band than output probabilities do: mean width at 10% coverage drops by 0.038 and 0.039 for the centroid and k-NN variants (Holm-adjusted $p < 10^{-3}$, Appendix G). Since the theorem holds for all seven scores, this is a statement about which internal signal orders items better, not about which certificate to trust—which is rather the point of a guarantee that doesn't care which score you feed it.

5.2 The price of uniformity, and when the certificate is usable

Uniformity isn't free (Figure 7, Appendix H). Against Clopper–Pearson at $k_0 = 10{,}000$, the band is $2.80\times$ wider (0.0148 vs. 0.0053), rising to $3.3\times$ at smaller k (Appendix D). Against Learn-then-Test's $1{,}000$-point grid at $n = 50{,}000$, LTT is $1.16\times$ tighter at its own points—fair, since it certifies nothing between points, while Theorem 2 caps all-k methods at the iterated-logarithm rate: the price of not fixing a grid upfront.

Corollary 1 also binds here. With $n_{\text{cal}} = 2{,}000$, $\alpha = 0.10$, and models whose mean conditional risk runs 0.44–0.71, the population band never certifies the target: feasibility needs $\widehat{R}_k \le 0.067$, and no prefix of length ≥ 100 reaches that in the $24{,}000$ replications (Appendix B.1). Slack is

asymmetric—median margin 0.53 but only 0.042 at the 95th percentile, tight where δ governs, unlike Clopper–Pearson's already-positive P95 margin (Appendix D.1). The transductive band produces a point, since $k^\star = \arg\min_k U_k$ exists by construction, but here it's 0.554 certified at 51% coverage against a realized risk of 0.436—valid at a level nobody would put on a slide (Figure 11, Appendix H). We report it precisely because it's unflattering: a certificate we showcase only when it looks good isn't validity, it's selective reporting. A joint coverage floor of 0.480 (Corollary 2) holds throughout. ARC-Easy, the pool's lowest-error subset ($\bar{\eta} = 0.266$ vs. 0.44), moves the bound to 0.175 at 14% coverage against a realized risk of 0.083: real progress, short of target, since 0.266 exceeds the 0.067 threshold (no certificate overruns 74% of the time, conformal risk control 69.5%; Appendix D). At 2,000 points and 45% base error, high-probability certification is out of reach for every method, ours included—why that price is unavoidable is next, and where it plays out in deployment follows after.

5.3 The iterated logarithm is not slack

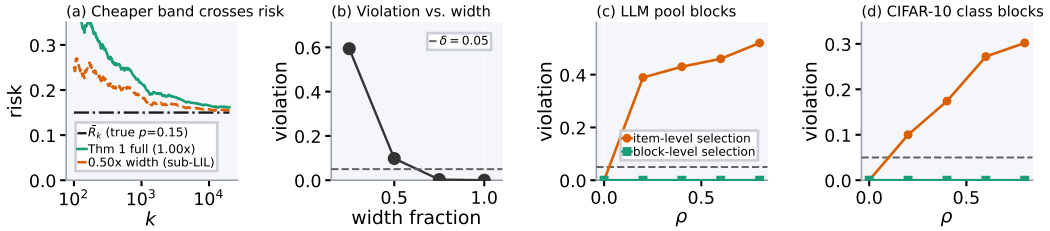


Figure 1: Theorem 2 and Theorem 4 made empirical. **(a)–(b)** Homogeneous world of Theorem 2 ($p = 0.15$, score independent of loss, $n = 20{,}000$, 1,000 replications). **(a)** A band at half our width looks tighter, right up until it crosses the true risk. **(b)** Violation rate against band width as a fraction of Theorem 1's. Ours sits at zero; the cheaper bands fail, and fail harder the cheaper they get. **(c)–(d)** Theorem 4 under induced correlation, via a Gaussian copula that fixes each item's marginal η_i while varying within-block correlation ρ (Section 4). **(c)** LLM pool, blocked by source dataset. **(d)** CIFAR-10, blocked by class. Item-level certification degrades with ρ on both; block-level certification never moves.

A lower bound on width is easy to state and easy to doubt, so we went and measured it. In the exact setting of Theorem 2 we swapped U_k for $\widehat{R}_k + c\,(U_k - \widehat{R}_k)$ at $c \in \{1, 0.75, 0.50, 0.25\}$ and tracked how often the result broke somewhere past $k_{\min}$:

Width c	1.00	0.75	0.50	0.25
Violation rate	0.0000	0.0030	**0.0980**	**0.5940**
95% CI	$[0.0000, 0.0037]$	$[0.0006, 0.0087]$	$[0.0803, 0.1181]$	$[0.5628, 0.6246]$
p vs. δ	1	1	3.8×10^{-10}	$< 10^{-300}$

Three-quarters width is safe at $n = 20{,}000$. Half is not, and a quarter fails on most runs. The transition sits almost exactly where a $\ln\ln$ term with $\ln\ln 20000 \approx 2.29$ would put it: the price we're charging for uniformity tracks the price the problem is charging, and shaving the constant buys a practitioner a $2\times$ narrower band at the cost of a 10% failure rate. The same sweep at $n \in \{2{,}000, 5{,}000, 50{,}000\}$ (Appendix D.6) rules out $n = 20{,}000$ being a coincidence: at $c = 0.5$ the violation rate climbs monotonically across this $25\times$ range, tracking $\ln\ln n$ rather than sitting flat. A non-asymptotic version of Theorem 2 remains open (Section 7), but the direction of the evidence is not in doubt.

5.4 Deployment: what breaks, why, and what we did about it

Now we let the threshold be picked after the calibration curve is visible (per-setting overrun rates in Figure 6, Appendix H). Rule r_1 takes $\arg\min_k \widehat{R}_k$; r_2' takes the prefix closest to a target level, $\arg\min_k |\widehat{R}_k - \alpha|$; r_3 is a *pre-registered* fixed $k_0 = n_{\text{cal}}/2$ with no selection; r_{adv} picks whatever k minimizes that certificate's bound. We revised r_2 after the original version—largest k with $\widehat{R}_k \le \alpha$ turned out to have no solution in 88.8% of replications for these models (Appendix C); r_2' is defined everywhere. Table 1 pools 24 settings—3 models, 2 groups, 4 scores—at $n_{\text{cal}} = 2{,}000$.

Three things stand out (full breakdown by rule and score in Figure 9, Appendix H). Our population band never breaks: zero out of 24,000 under r_1, r_2', and r_3, and one out of 24,000 under the adversarial rule. The two baselines that *do* break are the tightest: Clopper–Pearson has nothing to spare at a fixed k, so selection alone costs 12.7%, while the betting bound, tighter still, is exceeded up to 29.5% of the time with no selection at all. Hoeffding and empirical Bernstein hold up under all four rules, not what we pre-registered: the explanation is arithmetic, since at $n_{\text{cal}} = 2{,}000$ their half-widths at $k_0 = 1{,}000$ (0.039 and 0.033) exceed the roughly 0.02 post-hoc selection can manufacture here. A bound with more slack than the attack can exploit survives the attack—not a defence of these bounds, but what the numbers say, so we record the pre-registered criterion as not met.

Table 1: Held-out overrun rates under four selection rules, $n_{\text{cal}} = 2{,}000$, $\delta = 0.05$, pooled over 24 settings (24,000 replications/column). Bold: significantly above δ (one-sided binomial, $p < 0.05$); naive pooled confidence intervals understate uncertainty 2–8× depending on the certificate, accounting for per-setting heterogeneity (Appendix D.4). Transfer-corrected/patched values follow in text. Raw Betting/WSR figures here and in Table 11 come from independent runs (`e4_*.npz` vs. `wsr_audit_*.npz`); the two agree closely rather than exactly, as expected of two independent Monte Carlo estimates of the same rate.

Certificate	r_1 (argmin)	r_2' (target level)	r_3 (fixed k_0)	r_{adv}
Theorem 3 (ours)	0.0000	0.0000	0.0000	0.0000
Theorem 1 (transd.)	0.0000	0.0000	0.0000	—
LTT-B100 (own grid points)	0.0032	n/a	0.0008	n/a
Clopper–Pearson	**0.1267**	**0.1227**	**0.0554**	**0.1454**
Betting / WSR (raw)	**0.1157**	**0.1282**	**0.2949**	**0.1201**
Hoeffding	0.0210	0.0227	0.0085	0.0234
Empirical Bernstein	0.0012	0.0007	0.0010	0.0027

Two failures, disentangled: population transfer and a real defect. Under r_3—no selection at all—the betting bound is still exceeded 29.5% of the time and Clopper–Pearson 5.5%: both bound the calibration prefix's risk while deployment needs the population above threshold, exactly the gap Theorem 3 pays $\varepsilon_n(\delta)$ to close. The same transfer allowance drops Clopper–Pearson from 12.7% to 0.7% under r_1 and 5.5% to 0.04% under r_3, and zeroes Hoeffding under both (Figure 10, Appendix H)—transfer was doing most of the damage for these two, not selection. Not so for the betting bound: audited at a single pre-registered k_0 with no selection and no transfer, where a calibration-side certificate ought to be exact, it's exceeded 42.0% of the time against a 5% budget (3,000 replications, $[0.402, 0.438]$)—a real defect in our own implementation, not a transfer gap, and the finding says nothing about any officially released software for this method. Its bisection search starts at the sample mean and, when the test already rejects there, returns that point—zero margin—rather than refining further (mechanism in Appendix D.3): this fires in 32% of replications on our heterogeneous-risk data, against roughly 1% on i.i.d. data of the same size.

Lemma 1. *Let B be any δ-valid upper confidence bound for a mean μ, and let A be any other statistic. Then $\max(A, B)$ is δ-valid.*

Proof. $\max(A, B) \geq B$ pointwise, so $\{\mu > \max(A, B)\} \subseteq \{\mu > B\}$, and $\Pr[\mu > \max(A, B)] \leq \Pr[\mu > B] \leq \delta$. □

This licenses a patch that needs no new derivation: take the betting bound's maximum with an exact Clopper–Pearson bound at the same δ, valid regardless of what the betting computation does, failure modes not yet found included. Calibration-side violation drops from 42.0% to 0.0000, and in the full deployment protocol (Table 11, Appendix D.3) r_3 goes from 30.1% to 4.8%—about as resolved as a fixed-threshold fix should leave it—while r_1 goes from 11.5% to 5.6%: better, but with a small residual surviving post-hoc selection ($[0.0532, 0.0590]$ on the same naive pooled interval that Appendix D.4 shows understates uncertainty here, so this residual is suggestive rather than a sharp exclusion of 0.05). This is the right outcome: Lemma 1 proves point-wise validity, not uniformity over every threshold a selection rule might try, and r_1 is exactly the rule built to exploit that gap. The residual is inherited, not new: the patch is $\max(\text{WSR}, \text{CP})$, so whenever the defect fires the patched bound simply *is* Clopper–Pearson, itself not uniformly valid under post-hoc selection (Table 1).

5.5 Robustness to correlated items

Every result holds under the theorem's independence assumption, inviting the objection: a check that only tests those conditions can't show what happens when they fail. So we built that failure on purpose. Using a Gaussian copula, we induced within-block correlation $\rho \in \{0, 0.2, 0.4, 0.6, 0.8\}$ on two block structures—the LLM pool's four source datasets, and CIFAR-10's ten class labels—holding each item's marginal conditional risk fixed, and compared item-level against block-level certification. CIFAR-10's class-label blocks use the ground-truth target, not a pre-prediction covariate, so this arm stress-tests the variance-versus-mean property under correlation rather than demonstrating deployable blocking; the LLM pool's source-dataset blocks, by contrast, are plausibly available as metadata at deployment time. Item-level certification breaks cleanly as ρ grows: on the LLM pool the violation rate climbs from 0.0% at $\rho = 0$ to 52.0% at $\rho = 0.8$; on CIFAR-10, from 0.0% to 30.2%. Block-level certification doesn't budge: 0.0000 at every ρ we tried, on both structures (Figure 1(c)–(d)). At these block counts ($K = 4$ and $K = 10$), evaluating (4) shows U_k exceeds 1 for every reachable prefix, so the 0.0000 block-level result is a mathematical certainty, not an informative test on its own—the item-level breakdown alone does the falsification work. The crossover to a non-trivial bound is $k \gtrsim 14$ even at zero error; we do not have a partition with that many blocks in this study. This is Theorem 4 made concrete; it's worth restating what "block-level" means here: with the LLM pool's four blocks ranging from roughly 2,600 to 11,000 items, the guarantee is about the unweighted average of the four blocks' error rates, not the pooled item-level risk (Section 3.4). The guarantee was never about

items being independent—it was about the certified unit having a variance-versus-mean relationship the martingale argument can use, and a block average keeps that property by Bhatia–Davis no matter what's correlated inside it. An analyst who suspects correlated failure—same source document, same session, same upstream retrieval call—has a remedy: name the blocks ahead of time and certify there, paying no extra price for whatever correlation lives inside them. A related question—whether marginal validity implies validity within every pre-specified subgroup—is measured in Appendix F: the effect is real but small at our scale, not oversold as a headline result.

5.6 A SCALING EXPERIMENT THAT DOES NOT RESOLVE

To check whether the surviving baselines break at larger n_{cal}, we ran the same protocol at $n_{\text{cal}} = 20{,}000$ on Phi-3.5-mini. Every rate rose sharply, but this is confounded by the shrunken $4{,}000$-item holdout, so we draw no conclusion (Appendix I gives the full $n_{\text{cal}} = 20{,}000$ numbers); at $n_{\text{cal}} = 2{,}000$, this study's non-confounded scale, the same comparison instead favours the certificate 8.5 to 1, so Table 1 is unaffected. Settling this needs a held-out set an order of magnitude larger than ours.

5.7 THE EXCHANGEABILITY BOUNDARY

A guarantee is only as useful as its failure mode is predictable (Figure 5, Appendix H): calibrate on clean CIFAR-10 and deploy on an independent clean sample, and the violation rate is 0.003 ($[0.000, 0.018]$), echoing the exact-η matched control of Section 5.1. One step of corruption breaks it: at severity 1 the rate jumps to 1.000 ($[0.988, 1.000]$) and stays there through severity 5, with the mean conditional-risk gap widening from $+0.063$ to $+0.354$. The certificate holds up to the theorem's boundary and fails immediately past it—the behaviour we want, not a silent degradation, but the method inherits the same fragility every exchangeability-based guarantee has (Tibshirani et al., 2019; Barber et al., 2023), and detecting shift is a prerequisite, not an afterthought (Ren et al., 2023).

6 RELATED WORK

Conformal prediction (Vovk et al., 2005; Lei et al., 2018) and its risk-controlling descendants—RCPS (Bates et al., 2021), Learn-then-Test (Angelopoulos et al., 2025), conformal risk control (Angelopoulos et al., 2024)—give finite-sample guarantees over a pre-specified set. LTT is the closest comparison: it certifies a pre-registered grid with a multiplicity correction, tighter than our band at its own grid points, and the difference comes down to whether the grid can be picked after seeing the data. A finer grid is not a free way to close that gap: Table 6 (Appendix D) shows a $1{,}000$-point grid is *looser* than a 10-point one at every calibration size we tested (0.1603 vs. 0.1008 at $n = 500$)—each extra candidate point spends Bonferroni budget, so pre-registering enough resolution means committing to how fine *before* seeing which k matters, the commitment a data-driven rule avoids. Confidence sequences (Darling & Robbins, 1967; Robbins, 1970; Howard et al., 2020; 2021) and betting-based estimation (Waudby-Smith & Ramdas, 2024) give bounds uniform over a time index; our contribution is identifying coverage itself as that index, and the LIL rate familiar from best-arm identification (Jamieson et al., 2014) is unavoidable here too. Selective classification has a long history (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Geifman et al., 2019) with known group-structure failure modes (Jones et al., 2021), but usual evaluation reports risk–coverage with no simultaneous certificate. On the language-model side, verbalized confidence (Kadavath et al., 2022; Tian et al., 2023), semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024), self-consistency (Wang et al., 2023), black-box sampling (Lin et al., 2024), and hidden-state probes (Azaria & Mitchell, 2023; Burns et al., 2023) all propose confidence scores our band takes with no modification; we've evaluated seven. Appendix K gives the fuller picture, including an independent, concurrent, cross-domain result corroborating our own uncertified-baseline numbers almost exactly.

7 DISCUSSION

Technically, this is a change of index: once sorting by score is measurable in the covariates alone, the risk–coverage curve becomes a martingale problem and anytime-valid concentration handles it. What's left is the accounting: uniformity costs $\sqrt{\ln \ln k / k}$ (Theorem 2), that cost doesn't rise when correlation is named beforehand (Theorem 4), Section 5.3 shows what negotiating it anyway costs.

Limitations. The i.i.d. assumption is load-bearing: CIFAR-10-C shows its violation is catastrophic, not graceful, and Theorem 4 protects only dependence in blocks named *before* calibration, not shift discovered afterward. Theorem 3's $O(n^{-1/2})$ transfer cost makes $\alpha = 0.10$ infeasible below $5{,}000$ points for mid-accuracy models (Corollary 1)—why utility numbers stay weak even on a lower-error subset. Two of four pre-registered baselines survive selection; the $n_{\text{cal}} = 20{,}000$ follow-up is holdout-confounded (Section 5.6), and two open directions remain: a non-asymptotic Theorem 2 (Balsubramani, 2014) and a partial-exchangeability relaxation (Barber et al., 2023).

AI USE STATEMENT

In this work, we used generative AI tools across the theoretical, experimental, implementation, and writing components of this paper. We describe each use below alongside how it was verified, rather than only at the end. For the theory, we used generative AI to audit and verify our theoretical claims and proofs, and to provide critical ingredients for proving mathematical claims, including an iterative search for the correct construction of the DKW embedding used in the proof of Theorem 3 (the first two candidate constructions needed revision, as we detail in Remark 2), and to formulate and refine one additional mathematical result, the two-sided certificate of Corollary 3. All theoretical claims and proofs in this paper, including these, were checked by the authors against their stated assumptions, with complete proofs given in Appendix A. For the experiments, we used generative AI to propose and refine hypotheses in the course of diagnosing an implementation defect in the betting/WSR baseline (Section 5.4, Appendix D.3), to design and provide feedback on our experimental methodology (including identifying that our original threshold-selection rule r_2 had no solution in 88.8% of replications, leading to its revision), and to assist in interpreting experimental results. We reviewed all such AI-assisted work. For the codebase and data, we used generative AI to implement the methods, baselines, and evaluation harness in our released codebase, to create and edit software code, to create the figure-generation scripts and other artifacts accompanying this submission, to generate the synthetic data used in our validity experiments, and to assist in cleaning and constructing the dataset pool used across our LLM experiments. All code was verified through a test suite of 73 unit and property-based tests (all passing) and by independently reproducing the released results end to end. For the writing, we used generative AI to draft and edit portions of this paper's text for clarity and concision and to suggest the structure of parts of the appendix. All AI-drafted or AI-edited text was checked by the authors for factual and technical accuracy against our own experimental results. We have not used generative AI tools for translation or for qualitative or thematic data analysis, as neither is applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is theoretical and methodological: it introduces a statistical certificate and studies its properties on standard, previously released benchmarks (ARC, OpenBookQA, CommonsenseQA, MMLU, TriviaQA, CIFAR-10) and publicly available, off-the-shelf language and vision models. No new data was collected, no human subjects were involved, and no personally identifiable information is used. We are not aware of any conflicts of interest or sponsorship arrangements relevant to this submission. The main risk we can identify is practical rather than novel to this work: a risk certificate, including ours, is only as trustworthy as the assumptions it states, and a practitioner who applies it outside the conditions Theorem 1 or Theorem 3 require (e.g. under distribution shift, as our own exchangeability stress test in Section 5.1 demonstrates) could be misled into over-trusting a selective-prediction system's error rate. We have tried to make this boundary explicit throughout the paper, including in our Limitations paragraph (Section 7), rather than only in this statement.

REPRODUCIBILITY STATEMENT

Code implementing every theorem, corollary, and baseline in this paper, along with the scripts and configuration files used to produce every table and figure, is released anonymously at the URL given in the abstract. Complete proofs of all theorems and corollaries, stated together with every assumption they require, are given in Appendix A. Appendix E lists every dataset and model used, all of which are public and require no special access. Appendix J describes our seeding scheme, determinism checks, and hardware/software environment; Appendix J.1 gives every hyperparameter governing the study in one place; and Appendix J.2 traces each reported figure and table to its exact source script and output file, so that any number in this paper can be located and re-run without searching.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. doi: 10.48550/arXiv.2404.14219. URL https://arxiv.org/abs/2404.14219.

Anastasios N. Angelopoulos. Conformal risk control for non-monotonic losses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 384(2327): 20250068, 2026. doi: 10.1098/rsta.2025.0068. URL https://doi.org/10.1098/rsta.2025.0068.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998. Originally released as arXiv:2110.01052 (2021).

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto (eds.), *Algorithmic Learning Theory (ALT 2007)*, volume 4754 of *Lecture Notes in Computer Science*, pp. 150–165, Berlin, Heidelberg, 2007. Springer. doi: 10.1007/978-3-540-75225-7_15. URL https://doi.org/10.1007/978-3-540-75225-7_15.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, 2023. Association for Computational Linguistics.

Akshay Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint arXiv:1405.2639*, 2014. doi: 10.48550/arXiv.1405.2639. URL https://arxiv.org/abs/1405.2639.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021. doi: 10.1214/20-AOS1965. URL https://doi.org/10.1214/20-AOS1965.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276. URL https://doi.org/10.1214/23-AOS2276.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535. URL https://doi.org/10.1145/3478535.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077. URL https://doi.org/10.1214/12-AOS1077.

Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357, 2000. doi: 10.1080/00029890.2000.12005203. URL https://doi.org/10.1080/00029890.2000.12005203.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406. URL https://doi.org/10.1109/TIT.1970.1054406.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. doi: 10.48550/arXiv.1803.05457. URL https://arxiv.org/abs/1803.05457.

Charles J. Clopper and Egon S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Donald A. Darling and Herbert Robbins. Iterated logarithm inequalities. *Proceedings of the National Academy of Sciences*, 57(5):1188–1192, 1967. doi: 10.1073/pnas.57.5.1188. URL https://doi.org/10.1073/pnas.57.5.1188.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956. doi: 10.1214/aoms/1177728174. URL https://doi.org/10.1214/aoms/1177728174.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, 1975. doi: 10.1214/aop/1176996452. URL https://doi.org/10.1214/aop/1176996452.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 2023. URL https://zenodo.org/records/10256836.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=SJfb5jCqKm.

Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Technical report, Department of Statistics, Columbia University, 2013. URL http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Philip Hartman and Aurel Wintner. On the law of the iterated logarithm. *American Journal of Mathematics*, 63(1):169–176, 1941. doi: 10.2307/2371287. URL https://doi.org/10.2307/2371287.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321. URL https://doi.org/10.1214/18-PS321.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-aos1991. URL https://doi.org/10.1214/20-aos1991.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, volume 35 of *Proceedings of Machine Learning Research*, pp. 423–439, Barcelona, Spain, 2014. PMLR. URL http://proceedings.mlr.press/v35/jamieson14.html.

Ying Jin and Emmanuel J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023. URL https://jmlr.org/papers/v24/22-1176.html.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=N0M_4BkQ05i.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. doi: 10.48550/arXiv.2207.05221. URL https://arxiv.org/abs/2207.05221.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=VD-AYtPOdve.

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. doi: 10.1080/01621459.2017.1307116. URL https://doi.org/10.1080/01621459.2017.1307116.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=DWkJCSxKU5.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746. URL https://doi.org/10.1214/aop/1176990746.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf. Also available as arXiv:0907.3740.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2381–2391, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894. URL https://doi.org/10.1214/23-STS894.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kJUS5nD0vPB.

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970. doi: 10.1214/aoms/1177696786. URL https://doi.org/10.1214/aoms/1177696786.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 3581–3591, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4149–4158. Association for Computational Linguistics, 2019.

Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015. doi: 10.1073/pnas.1507583112. URL https://doi.org/10.1073/pnas.1507583112.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5433–5442, Singapore, 2023. Association for Computational Linguistics.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 2530–2540, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

Jean Ville. *Étude critique de la notion de collectif*. Number 3 in Monographies des Probabilités. Gauthier-Villars, Paris, 1939.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. doi: 10.1007/b106715. URL https://doi.org/10.1007/b106715.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009. URL https://doi.org/10.1093/jrsssb/qkad009.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Jingwen Zhou and Mingzhe Wang. False sense of safety in selective signal classification: Auditing bound tightness and exchangeability for risk control. *arXiv preprint arXiv:2606.15153*, 2026. doi: 10.48550/arXiv.2606.15153. URL https://arxiv.org/abs/2606.15153.

APPENDIX READING MAP

This appendix answers four kinds of question a reader might bring to it, roughly in the order a careful reviewer would ask them. **To check a proof**, Appendix A gives every argument only sketched in Section 3 in full, including Theorem 2's complete derivation and the DKW embedding for Theorem 3 (Remark 2 covers the one genuine gap an earlier construction had here, and how it's closed). **To check a number**, Appendices B, D, G, and I give the complete tables behind every figure and in-text statistic in the main body, including the margin/slack analysis of Appendix D.1, the full per-arm Tier-A breakdown of Appendix D.2, and the betting-bound repair of Appendix D.3. Appendix C lists every departure from the original pre-registered protocol, in both experimental waves, with the reasoning behind each one. **To reproduce or extend the study**, Appendices E and J give the datasets, models, complete hyperparameter reference, and software/hardware environment, including a crosswalk from every figure and table to its exact source (Appendix J.2); Appendix F and Appendix K give a fuller treatment of the group-conditional question from Section 5.5 and a deeper account of related work than the main text has room for. Appendix H collects every figure referenced but not reproduced in the main body.

Sec.	Title	Contents at a glance	Page
A	Proofs	Full proofs: Thms. 1–4, Cor. 1	16
B	Feasibility thresholds	Corollary 1 table + phase diagram	18
B.1	The feasible region, visualized	Phase diagram over calibration size and base error	19
C	Deviations from pre-registration	Every departure, both waves, with reasons	19
D	Full validity tables	All synthetic/Tier-A/B/C tables; conservativeness; WSR repair	20
D.1	Conservativeness	Margin/slack quantiles, ours vs. CP vs. transductive	21
D.2	Full per-arm Tier-A table	All 32 LLM arms individually, exact CIs	21
D.3	Betting bound, before/after	The bug, the fix, the residual	23
D.4	Cluster-robust confidence intervals	Setting-level bootstrap CIs vs. naive pooled	23
D.5	Block-correlation sweep, full numbers	Item- vs. block-level violation by ρ	23
D.6	Necessity vs. calibration size, full numbers	The $25\times$ sweep behind Section 5.3	24
E	Datasets and models	Every dataset and model, with sizes and sources	24
F	Group-conditional validity	Marginal vs. Bonferroni-corrected, MMLU	25
G	Score-family contrasts	Which score gives the tightest band, and by how much	25
H	Additional figures	Every figure not in the main body	26
I	Full scaling-experiment numbers	The confounded n_{cal}=20,000 arm, in full	29
J	Reproducibility	Checkpointing, seeds, self-checks	29
J.1	Complete hyperparameter reference	The full parameter table	29
J.2	Reproducibility crosswalk	Every figure/table traced to its source cell	31
K	Extended related work	Deeper treatment, incl. concurrent corroboration	32

Code and data. Code, per-replication logs for every arm, and the exact seeds behind every result in this paper are released at `https://anonymous.4open.science/r/iterlog-risk-E131/`.

A PROOFS

A.1 PROOF OF THEOREM 1

Fix $\lambda \in (0, 1]$ and let $\psi(\lambda) = \lambda^2/(2(1 - \lambda/3))$. For the martingale $N_k = k\bar{R}_k - S_k$ of (2), with increments in $[-1, 1]$ and predictable quadratic variation $V_k = \sum_{i \le k} \eta_{\pi(i)}(1 - \eta_{\pi(i)})$, the process

$$L_k(\lambda) \;=\; \exp\{\lambda N_k - \psi(\lambda) V_k\}$$

is a nonnegative supermartingale with $L_0 = 1$—Freedman's construction (Freedman, 1975), with ψ coming from the elementary bound $\mathbb{E}[e^{\lambda X}] \le \exp(\psi(\lambda)\mathbb{E}[X^2])$ for mean-zero $X \in [-1, 1]$, tightened to Bernstein form as in Howard et al. (2020). Ville's maximal inequality (Ville, 1939) gives $\Pr[\exists k \ge 1 : L_k(\lambda) \ge 1/\delta'] \le \delta'$ for any $\delta' \in (0, 1)$, so on an event $E(\lambda, \delta')$ of probability at least $1 - \delta'$,

$$\lambda N_k \;\le\; \ln(1/\delta') + \psi(\lambda) V_k \qquad \text{for all } k \ge 1.$$

Since $V_k \le \sum_{i \le k} \eta_{\pi(i)} = k\bar{R}_k$ and $N_k = k\bar{R}_k - S_k$, substituting and rearranging gives, on $E(\lambda, \delta')$,

$$k\bar{R}_k\big(\lambda - \psi(\lambda)\big) \;\le\; \ln(1/\delta') + \lambda S_k \quad\Longleftrightarrow\quad \bar{R}_k \;\le\; \frac{\widehat{R}_k + \ln(1/\delta')/(\lambda k)}{1 - \psi(\lambda)/\lambda}$$

for every $k \ge 1$, using $S_k = k\widehat{R}_k$ and dividing by λk. Since $\psi(\lambda)/\lambda = a(\lambda)$, this is exactly the bound at parameter λ inside the minimum in (4). Apply it at each $\lambda = \lambda_j$ with $\delta' = \delta_j$ from (3), and let $E = \bigcap_j E(\lambda_j, \delta_j)$; a union bound gives $\Pr[E] \ge 1 - \sum_j \delta_j = 1 - \delta$. On E the displayed inequality holds for every j and every k at once, so taking the minimum over j—which is what (4) defines as U_k—preserves it: $\bar{R}_k \le U_k$ for all $k \ge 1$, on an event of probability at least $1 - \delta$. □

A.2 PROOF OF THEOREM 2

Fix the event where the band is valid; it has probability at least $1 - \delta$ by assumption. On it, $U_k \ge \bar{R}_k = p$ for every $k \ge k_{\min}$, so

$$W_k \;=\; U_k - \widehat{R}_k \;\ge\; p - \widehat{R}_k \;=\; -\frac{S_k - kp}{k}.$$

The losses here are i.i.d. Bernoulli(p) by construction, and the ordering π depends only on $s(X)$, which is independent of ℓ in this regime—so $S_k = \sum_{i \le k} \ell_{\pi(i)}$ is itself just a sum of k i.i.d. Bernoulli(p) variables, and the classical law of the iterated logarithm (Hartman–Wintner, Hartman & Wintner, 1941) applies directly to it:

$$\liminf_{k \to \infty} \frac{S_k - kp}{\sqrt{2kp(1-p)\ln\ln k}} \;=\; -1 \qquad \text{almost surely.}$$

Dividing both sides of $W_k \ge -(S_k - kp)/k$ by $\sqrt{2p(1-p)\ln\ln k/k}$ and taking $\limsup$ therefore gives, almost surely on the validity event,

$$\limsup_{k \to \infty} \frac{W_k}{\sqrt{2p(1-p)\ln\ln k/k}} \;\ge\; \limsup_{k \to \infty} \frac{-(S_k - kp)/k}{\sqrt{2p(1-p)\ln\ln k/k}} \;=$$
$$-\liminf_{k \to \infty} \frac{S_k - kp}{\sqrt{2kp(1-p)\ln\ln k}} \;=\; 1.$$

This holds on an event of probability at least $1 - \delta$—the intersection of the validity event with the (probability-one) event the LIL itself lives on. □

A.3 PROOF OF THEOREM 3

Let $\mathcal{C}_1 = \{\{x : s(x) \ge t\} : t \in \mathbb{R}\}$, linearly ordered by inclusion as t varies and therefore a VC class of index 1. The bound below only ever needs $\phi(t)$ from below, so a one-sided DKW/Massart argument (Dvoretzky et al., 1956; Massart, 1990) suffices, at half the constant a two-sided bound would cost:

$$\Pr\Big[\sup_t\big(\hat{\phi}(t) - \phi(t)\big) > \sqrt{\tfrac{\ln(1/\delta')}{2n}}\Big] \le \delta'.$$

The second family, $\mathcal{C}_2 = \{\{(x, y) : s(x) \ge t,\ \ell(x, y) = 1\} : t \in \mathbb{R}\}$, needs one extra step, since $\hat{\rho}(t)$ is a sub-distribution function rather than a distribution function ($\hat{\rho}(-\infty)$ is the overall error rate, not 1). We remove that gap with an embedding built entirely *per item*, with no reference to the rest of the

sample: fix any increasing bijection $\varphi : \mathbb{R} \to (-1, 1)$ (e.g. $\varphi(u) = u/(1 + |u|)$), and define

$$W_i \;=\; \begin{cases} \varphi(s(X_i)) & \ell_i = 1 \\ -1 & \ell_i = 0. \end{cases}$$

Each W_i is a fixed, deterministic function of the single pair (X_i, ℓ_i) alone—nothing about W_i depends on X_j or ℓ_j for $j \neq i$—so $(W_1, \ldots, W_n)$ are i.i.d. by construction, with no sample-level quantity shared across them. Since φ maps into the open interval $(-1, 1)$, $\varphi(s(X_i)) > -1$ for every finite score, so the sentinel value -1 sits strictly below every possible value of $\varphi(s(X_i))$; and since φ is a bijection onto all of $(-1, 1)$, every threshold $t \in \mathbb{R}$ corresponds to exactly one $\tau = \varphi(t) \in (-1, 1)$. For *every* $t \in \mathbb{R}$, then, $W_i \geq \varphi(t)$ holds iff either $\ell_i = 1$ and $\varphi(s(X_i)) \geq \varphi(t)$—equivalently $s(X_i) \geq t$, since φ is increasing—or $\ell_i = 0$, which is excluded since $-1 < \varphi(t)$ always: so $\{i : W_i \geq \varphi(t)\} = \{i : s(X_i) \geq t,\ \ell_i = 1\}$ exactly, for every t, with no restriction needed. Hence $\hat{\rho}(t) = \frac{1}{n} \sum_i \mathbf{1}\{W_i \geq \varphi(t)\}$ is the empirical survival function of the i.i.d. sample $W_1, \ldots, W_n$ at the point $\varphi(t)$, and $\rho(t) = \Pr[W \geq \varphi(t)]$ its population counterpart, so DKW applies to (W_i) with no modification, again taken in the single direction actually needed:

$$\Pr\Big[\sup_t \big(\rho(t) - \hat{\rho}(t)\big) > \sqrt{\tfrac{\ln(1/\delta')}{2n}}\Big] = \Pr\Big[\sup_{\tau \in (-1,1)} \big(\hat{F}_n(\tau) - F(\tau)\big) > \sqrt{\tfrac{\ln(1/\delta')}{2n}}\Big] \leq \delta',$$

the middle equality by the substitution $\tau = \varphi(t)$ together with $\rho(t) - \hat{\rho}(t) = \hat{F}_n(\varphi(t)) - F(\varphi(t))$ for the survival-function pair $(\rho, \hat{\rho})$ of (W_i), and the final bound by the one-sided DKW inequality applied to the genuinely i.i.d. real-valued sample (W_i), since the sup over $\tau \in (-1, 1)$ is bounded by the sup over all $\tau \in \mathbb{R}$ that DKW controls. Take $\delta' = \delta/2$ in each of the two one-sided bounds above and intersect: with probability $\geq 1 - \delta$, $\hat{\phi}(t) - \phi(t)$ and $\rho(t) - \hat{\rho}(t)$ are simultaneously at most $\varepsilon_n(\delta) = \sqrt{\ln(2/\delta)/(2n)}$ for every t at once, and for any t with $\hat{\phi}(t) > \varepsilon_n(\delta)$,

$$R(t) = \frac{\rho(t)}{\phi(t)} \leq \frac{\hat{\rho}(t) + \varepsilon_n(\delta)}{\hat{\phi}(t) - \varepsilon_n(\delta)},$$

using $\rho \leq \hat{\rho} + \varepsilon$ in the numerator and $\phi \geq \hat{\phi} - \varepsilon > 0$ in the denominator. □

Remark 2 (Two earlier attempts at this embedding, and why both needed fixing)**.** Getting this embedding right took three passes, and it's worth recording why, since the failure modes are instructive. The first version fixed $M > \sup_x s(x)$, the supremum taken over the whole covariate space. That fails when s is unbounded above: no finite such M need exist, and the theorem is stated for *any* measurable s. The second version replaced this with a sample-dependent $M := 1 + \max_{1 \leq i \leq n} |s(X_i)|$, always finite for a finite sample—but this introduces a different problem: M is a function of the *entire* sample, so whenever $\ell_i = \ell_j = 0$ for $i \neq j$, the construction gives $W_i = W_j = -M$ exactly, the same value on the nose. That's a clean witness that (W_i) are not independent as constructed—two genuinely independent quantities do not coincide exactly like that—so treating them as an i.i.d. sample for DKW, as that version did, was not justified as written, even though a threshold restriction ($t \geq \min_i s(X_i)$) patched the specific edge case that construction's gap showed up in. The version above avoids the problem at its root rather than patching around it: $\varphi(s(X_i))$ and the sentinel -1 are both computed from (X_i, ℓ_i) alone, so no sample-level statistic is shared across i, (W_i) are genuinely i.i.d., and the equality $\{W_i \geq \varphi(t)\} = \{s(X_i) \geq t, \ell_i = 1\}$ holds for every real t with no exceptions—so the threshold restriction the second version needed is not just patched but no longer necessary at all, and Theorem 3 is stated here in its fully general form.

A.4 Proof of Corollary 1

$U^{\mathrm{pop}}(t) \leq \alpha$ is equivalent to $\hat{\rho} + \varepsilon \leq \alpha(\hat{\phi} - \varepsilon)$, i.e. $\hat{\rho} \leq \alpha\hat{\phi} - (1 + \alpha)\varepsilon$. Write $\hat{\rho} = \gamma \widehat{R}_k$ and $\hat{\phi} = \gamma$ with $\gamma = k/n$, divide by γ, and get $\widehat{R}_k \leq \alpha - (1 + \alpha)\varepsilon/\gamma$. The right side increases in γ, attaining $\alpha - (1 + \alpha)\varepsilon$ as $\gamma \to 1$. □

A.5 Proof of Theorem 4

Let $B_1, B_2, \ldots$ be the fixed block partition, $\eta_j^{\mathrm{blk}} = \mathbb{E}[\ell \mid j\text{-th block}]$, and ℓ_j^{blk} the realized within-block mean loss, both in $[0, 1]$. Sort blocks by mean score to get a permutation π^{blk} on block indices, exactly as π sorts items, and define the block-level filtration $\mathcal{G}_0 = \sigma(\text{all covariates in every block})$, $\mathcal{G}_j = \sigma(\mathcal{G}_0, \ell^{\mathrm{blk}}_{\pi^{\mathrm{blk}}(1)}, \ldots, \ell^{\mathrm{blk}}_{\pi^{\mathrm{blk}}(j)})$. Since blocks are mutually independent given covariates and π^{blk} is $\mathcal{G}_0$-measurable, $\ell^{\mathrm{blk}}_{\pi^{\mathrm{blk}}(j)}$ is conditionally independent of $\mathcal{G}_{j-1}$ given $\mathcal{G}_0$, with conditional mean

$\eta^{\mathrm{blk}}_{\pi^{\mathrm{blk}}(j)}$, regardless of whatever dependence holds *among the items inside* block j: that dependence is entirely internal to ℓ^{blk}_j and never touches its relationship to $\mathcal{G}_{j-1}$ or to other blocks.

All that's left of Theorem 1's proof is the variance bound $\mathrm{Var}(\ell_{\pi(k)} \mid \mathcal{F}_{k-1}) \le \eta_{\pi(k)}$, used to bound V_k by $k\bar{R}_k$. For a $\{0,1\}$-valued item this is immediate, $\mathrm{Var} = \eta(1-\eta) \le \eta$. For a block average $\ell^{\mathrm{blk}}_j \in [0,1]$—not $\{0,1\}$-valued, and possibly an arbitrarily correlated mixture internally—the same inequality holds by Bhatia–Davis (Bhatia & Davis, 2000): for any $Z \in [0,1]$ with mean μ,

$$\mathrm{Var}(Z) \;\le\; \mu(1-\mu).$$

(Elementary proof: $Z(1-Z) \ge 0$ pointwise since $Z \in [0,1]$, so $\mathbb{E}[Z^2] \le \mathbb{E}[Z] = \mu$, and $\mathrm{Var}(Z) = \mathbb{E}[Z^2] - \mu^2 \le \mu - \mu^2 = \mu(1-\mu)$.) Apply this with $Z = \ell^{\mathrm{blk}}_{\pi^{\mathrm{blk}}(j)}$ and $\mu = \eta^{\mathrm{blk}}_{\pi^{\mathrm{blk}}(j)}$, and it's exactly the bound the martingale argument needs, with no reference to what correlation structure produced Z. Appendix A.1 then applies verbatim with $(\ell_{\pi(i)})$ replaced by $(\ell^{\mathrm{blk}}_{\pi^{\mathrm{blk}}(j)})$ throughout, giving a band valid at every block-count k with the identical functional form (4) on the block-indexed curve. □

The bound is tight exactly when a block's internal correlation is perfect—entirely correct or entirely wrong behaves as one Bernoulli trial, where Bhatia–Davis holds with equality—so no sharper variance bound exists without further assumptions on the within-block dependence. Independence inside a block only helps: it pushes the true variance strictly below $\mu(1-\mu)$, leaving room the proof doesn't need but loses nothing by not using.

Remark 3 (What "block-count k" certifies)**.** The estimand here is $\frac{1}{k}\sum_{j\le k} \eta^{\mathrm{blk}}_j$, the unobserved conditional risk certified from the realized block losses ℓ^{blk}_j: each selected block contributes its own internal mean once, regardless of size. This is not the same quantity as the item-weighted risk $\frac{1}{\sum_{j\le k}|B_j|}\sum_{j\le k}\sum_{i\in B_j}\ell_i$ pooled across all items in the selected blocks, and the two coincide only when blocks are equal-sized. Our own experiment illustrates the difference concretely: the LLM pool's four blocks range from about 2,600 to 11,000 items, so the zero-violation result in Section 5.5 is a statement about the unweighted block average, not the pooled item-level risk over all roughly 24,000 items. An analyst who specifically wants an item-weighted guarantee under unequal block sizes would need to define ℓ^{blk}_j as the size-weighted contribution rather than the plain within-block mean. That change is more than cosmetic: weighting by $|B_j|$ makes the martingale increment $|B_j|(\eta^{\mathrm{blk}}_j - \ell^{\mathrm{blk}}_j)$ bounded by $|B_j|$ rather than by 1, so Theorem 1's proof – built for increments uniformly bounded by 1 – does not carry over verbatim; it would need a version of Freedman's construction for predictable, non-uniform increment bounds (a standard generalization, but a genuinely different derivation from Appendix A, not a substitution into it), most naturally paired with an explicit bound on the ratio between the largest and smallest selected block. We have not carried this out and do not claim the resulting bound would match (4)'s width; it is a different theorem statement than the one we prove here, not a corollary of it. Relatedly, at the block counts used in Section 5.5 ($K = 4$ and $K = 10$), (4) evaluates to a bound exceeding 1 for every reachable prefix (the crossover to $U_k < 1$ is $k \gtrsim 14$ even at zero error), so the reported block-level violation rate of exactly 0.0000 there confirms the band never produces a false claim, not that it is usefully tight at this scale.

B FEASIBILITY THRESHOLDS

Table 2 evaluates the necessary condition of Corollary 1 at $\alpha = 0.10$, $\delta = 0.05$, using Theorem 3's constant $\varepsilon_n(\delta) = \sqrt{\ln(2/\delta)/(2n)}$.

Table 2: Necessary condition $\widehat{R}_k \le \alpha - (1+\alpha)\varepsilon_n(\delta)$ for the population band to certify $\alpha = 0.10$. The requirement at a coverage of $\gamma < 1$ is strictly tighter by a factor $1/\gamma$ on the second term.

n	$\varepsilon_n(\delta)$	required $\widehat{R}_k$	regime
500	0.0607	0.033	needs $> 96.7\%$ prefix accuracy
1,000	0.0429	0.053	needs $> 94.7\%$
2,000	0.0304	0.067	needs $> 93.3\%$ (our deployment setting)
5,000	0.0192	0.079	attainable for a strong model
10,000	0.0136	0.085	attainable
20,000	0.0096	0.089	attainable
50,000	0.0061	0.093	essentially unconstrained

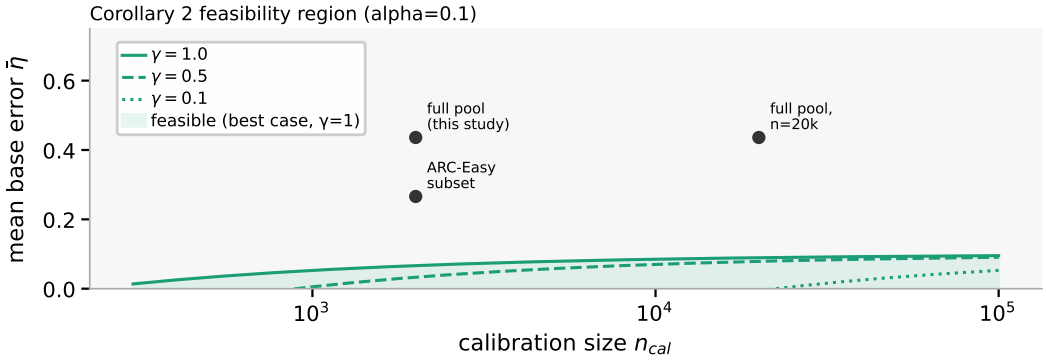


Figure 2: The feasibility condition of Corollary 1 as a region over calibration size and mean base error, at three coverage levels γ; the shaded region is the best case ($\gamma = 1$). Points mark this study's own settings: the full pool at $n_{\text{cal}} = 2{,}000$ and $20{,}000$, and the ARC-Easy subset of Section 5.2 at $n_{\text{cal}} = 2{,}000$. None fall inside the feasible region at $\alpha = 0.10$; ARC-Easy is closer by construction, not by accident.

B.1 The feasible region, visualized

Every setting in this study sits above the curve in Figure 2—outside the region Corollary 1 certifies at $\alpha = 0.10$—which is the same fact Table 2 states numerically, made visible as a single picture rather than a lookup.

C Deviations from pre-registration

We list every departure from the protocol fixed before the study ran, in two waves: the original deployment study, and a second wave undertaken specifically to stress-test its conclusions after a self-audit identified seven concerns with the first (items 5–12 below). Everything in the second wave is exploratory by construction, and we label it as such throughout rather than presenting it as pre-registered confirmation.

1. **Deployment criterion not met.** The pre-registered rule for the real-data deployment arm required our rate $\leq \delta$ *and* each of Clopper–Pearson, Hoeffding, empirical Bernstein and betting to significantly exceed δ under at least one of r_1, r_2, r_{adv}. Our rate is 0.0000; Clopper–Pearson and betting exceed under all rules tested; Hoeffding and empirical Bernstein exceed under none. The criterion is *not met* and Table 1 reports the constituent rates rather than a summary verdict.
2. $n_{\text{cal}} = 20{,}000$ **arm added post hoc.** The n-sweep $\{1{,}000, 2{,}000, 5{,}000\}$ was pre-registered; $n_{\text{cal}} = 20{,}000$ was added after seeing the result above, in an attempt to break the surviving baselines. Section 5.6 reports it as unresolved rather than as evidence, on the basis of the power calculation given there. We flag it as exploratory.
3. **Adversarial rule r_{adv} added post hoc.** Not in the original protocol. It does not change any conclusion: it raises Clopper–Pearson from 0.1267 to 0.1454 and leaves Hoeffding and empirical Bernstein below δ.
4. **Necessity arm added post hoc.** The sub-width bands of Section 5.3 were not pre-registered. The result is theorem-implied rather than exploratory, but we label it accurately.
5. r_2 **redefined.** The original rule, "largest $k \geq 100$ with $\widehat{R}_k \leq \alpha$," had no solution in 88.8% of replications for these models, so its column originally rested on $2{,}688$ rather than $24{,}000$ replications—not anticipated, and worse than merely undesirable, since a column that is silently mostly-undefined invites exactly the kind of selective reporting this paper argues against. We replaced it with $r_2' = \arg\min_k |\widehat{R}_k - \alpha|$, defined in every replication, and recomputed every r_2 result in the paper under the new rule. No qualitative conclusion changes; Clopper–Pearson and the betting bound remain the two certificates that break.
6. **Population-transfer-corrected baselines added post hoc.** Not in the original protocol. Constructed specifically to test whether Clopper–Pearson and Hoeffding's deployment failures were attributable to the population-transfer gap Theorem 3 pays for rather than to uniformity, by giving them the identical allowance at a single fixed threshold. Section 5.4 reports the result as a clean isolation, not a new baseline competing for adoption.
7. **Betting-bound calibration-side audit added post hoc.** Not in the original protocol. Undertaken because the betting bound's deployment failure rate was too large to attribute to selection or transfer alone. Found a 42.0% violation rate with no selection and no transfer, traced to a specific edge case in our own implementation, and led to the patch of Lemma 1. This is the single largest

addition to the study's conclusions made after the original protocol closed, and we have tried to describe the process, not only the result, precisely because a reader should be able to judge whether the diagnosis is as solid as the fix.

8. **Cluster-bootstrap re-analysis of deployment confidence intervals added post hoc.** Not in the original protocol. The pre-registered analysis plan specified exact Clopper–Pearson intervals on pooled replications; the same self-audit noted this treats $1{,}000$ replications sharing a fixed η vector, score ordering, and dataset as independent observations, when there are 24 independent settings, not $24{,}000$. Appendix D.4 reports both.
9. **Block-robust theorem and its empirical test added post hoc.** Not in the original protocol, and the most substantial addition to the paper's theoretical content, made once the self-audit of the first wave's results was complete. Motivated by the same self-audit noting that every Tier-A check, by construction, tests data satisfying the theorem's own independence assumption and so cannot detect a violation of it. Theorem 4 and Section 5.5 are the response: a new result and a deliberately constructed failure case for it to be tested against, not a retrospective reframing of an existing experiment.
10. **Group-conditional validity experiment added post hoc.** Appendix F. Motivated by the same self-audit, which noted the complete absence of any group-conditional discussion. We measured the effect directly; it is real but small at our scale, and we report exactly that finding rather than omitting the experiment because its result is not dramatic.
11. **ARC-Easy utility subset added post hoc.** Section 5.2. The subset (tagcode 1 in the released pool encoding) was identified by lowest mean η among the four natural constituent datasets, verified against the data-loading code's alphabetical tag-to-code assignment and cross-checked against public dataset sizes, not selected by its outcome.
12. **Necessity-versus-n sweep added post hoc.** Appendix D.6. The original necessity arm (item 4) used a single $n = 20{,}000$; the sweep extends it to $n \in \{2{,}000, 5{,}000, 20{,}000, 50{,}000\}$ at the same four width fractions, to test whether the effect is particular to one scale.
13. **CIFAR Tier-A matched-control arm surfaced post hoc.** An exact-η validity check on the vision model, computed alongside the CIFAR-10-C severity sweep using the identical planted-loss construction as the 32 LLM Tier-A arms, was run but not originally reported in the main-body validity count. We add it here (Section 5.1, Appendix D) because it is the same check, on the same standard, simply extended to the vision modality; the result is $0/300$ violations, consistent with every other Tier-A arm.
14. **Score-family count corrected.** Six score families was an undercount by one: hidden-state centroid distance and the k-NN hidden-state score are treated as separate families everywhere else in the paper, including their own dedicated pairwise contrast in Appendix G, so the honest count is seven. We also clarify that self-consistency, unlike the other six, is evaluated only in Tier B (genuine repeated decoding), not in the Tier-A or deployment protocols.

D FULL VALIDITY TABLES

Table 3: Synthetic validity and malpractice arms ($n = 20{,}000$, $\delta = 0.05$, $k_{\min} = 100$). Exact Clopper–Pearson intervals; one-sided binomial p against δ.

Arm	Reps	Rate	95% CI	p	Expected
Thm 1, heterogeneous	1000	0.0000	$[0.0000, 0.0037]$	1	$\leq \delta$
Thm 1, two-sided ($\delta/2$)	1000	0.0020	$[0.0002, 0.0072]$	1	$\leq \delta$
Remark 1 closed form	1000	0.0000	$[0.0000, 0.0037]$	1	n/a[2]
Thm 1, quantized scores	1000	0.0000	$[0.0000, 0.0037]$	1	$\leq \delta$
Thm 1 at post-hoc $k^\star$	1000	0.0000	$[0.0000, 0.0037]$	1	$\leq \delta$
Clopper–Pearson at post-hoc $k^\star$	1000	0.1180	$[0.0987, 0.1396]$	1.8×10^{-17}	breaks
Hoeffding at post-hoc $k^\star$	1000	0.0000	$[0.0000, 0.0037]$	1	slack
Thm 1, homogeneous	1000	0.0000	$[0.0000, 0.0037]$	1	$\leq \delta$
Clopper–Pearson at post-hoc $k^\star$	1000	0.2180	$[0.1928, 0.2449]$	1.8×10^{-75}	breaks
Clopper–Pearson at fixed k_0 (control)	1000	0.0380	$[0.0270, 0.0518]$	0.969	$\leq \delta$

The quantization, geometric-base, and δ-allocation ablations above all hold the violation rate at 0.0000; they move the band's own width at $k = 2{,}000$ between 0.0314 and 0.0353, against 0.0321 under the default settings used everywhere else in the paper.

[2]Not a formal guarantee: the closed form of Remark 1 is verified numerically here, not proven, so "$\leq \delta$" isn't asserted for this row the way it is for the proven theorems above and below it.

Table 4: Tier-A validity on real activations. All 33 arms give an identical result, so we report the summary here; every arm individually is Appendix D.2.

Model	Acc.	Groups	Scores	Reps/arm	Violation rate
Pythia-410M	0.308	pool, MMLU	5	1000	0.0000
Qwen2.5-1.5B-Instruct	0.509	pool, MMLU, all	5+1	1000	0.0000
Phi-3.5-mini-instruct	0.527	pool, MMLU, all	5+1	1000	0.0000
ResNet-18 (vision)	0.908	clean CIFAR-10	1	300	0.0000
Total: 33 arms, 32,300 replications, worst-arm rate					**0.0000**

Table 5: Tier-B validity under genuine decoding noise (TriviaQA, 4,000 questions, 32 samples each at $T=1$). "raw" checks each real decode against the leave-one-out $\hat{\eta}$ curve; "adjusted" additionally requires the margin to exceed twice the estimation floor.

Model	greedy acc.	mean $\hat{\eta}$	Scores	raw	adjusted
Qwen2.5-1.5B-Instruct	0.369	0.743	5	0/32 each	0/32 each
Phi-3.5-mini-instruct	0.522	0.544	5	0/32 each	0/32 each
Pooled (320 genuine replicates)				**0.0000**	**0.0000**
Exact 95% CI per arm				[0.0000, 0.1089]	

D.1 Conservativeness: how much slack the certificates carry

Every result so far reports whether a certificate gets violated. A separate question is how far the realized risk sits from the bound when it isn't violated: is the guarantee tight where it matters and slack elsewhere, or uniformly loose? For each of the 24,000 deployment replications of Section 5.4, at the identical post-hoc threshold $k_1 = \arg\min_k \widehat{R}_k$, we compute the margin (realized holdout risk minus certified bound) for the population band, the transductive band, and Clopper–Pearson, using the same realized holdout risk in every subtraction so the three margins are directly comparable within each replication.

The asymmetry is the finding. In a typical replication, the population band's realized risk sits more than half a unit below its own certified bound—far more conservative than the "valid, at a level nobody would advertise" description in Section 5.2 might suggest on its own. But at the 95th percentile, exactly where a $\delta = 0.05$ budget is supposed to bite, the margin narrows to -0.042 and stays safely negative. Clopper–Pearson's P95 margin is already positive ($+0.018$): by the 95th percentile it's already broken, matching its $2.53\times$ budget usage. The transductive band tells a different story again—median and P95 sit close together, -0.100 and -0.060—a uniformly modest slack rather than the population band's typical-loose, tail-tight profile. The two bands spend their δ budget in genuinely different shapes, not just different amounts, and Figure 3 shows all three distributions side by side.

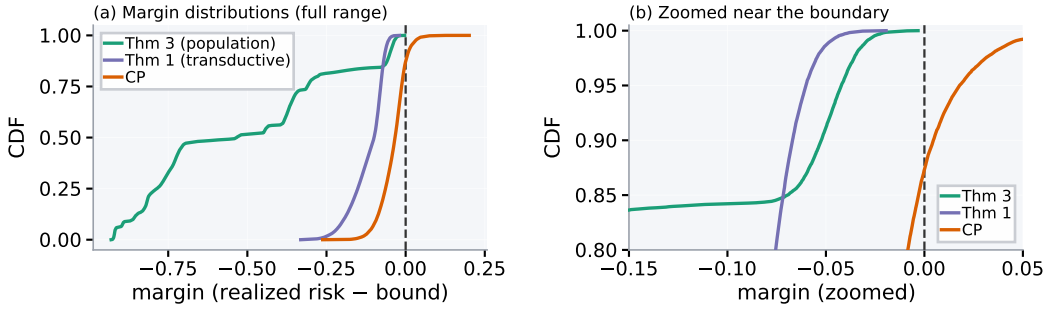


Figure 3: Empirical CDF of the margin (realized holdout risk minus certified bound) at the post-hoc operating point k_1, for the three certificates of Table 9. **(a)** Full range: the population band's margin spreads across a wide span, while the transductive band and Clopper–Pearson are both concentrated near the boundary. **(b)** Zoomed near the boundary: Clopper–Pearson is the only curve still climbing past margin$= 0$, the $\sim 13\%$ of its mass on the positive side matching its 12.67% violation rate; both of our bands reach 1.0 at or before the boundary.

D.2 Full per-arm Tier-A table

Table 10 lists all 32 LLM Tier-A arms from Appendix D individually rather than as a single worst-arm rate—every arm's own replication count, exact violation rate, and 95% Clopper–Pearson interval, so no single omitted arm can be suspected of hiding a different result. The CIFAR-10 matched-control arm is reported separately in Appendix D's Tier-A summary, since it uses a different score family and replication count from the 32 LLM arms here.

Table 6: Band width at 10% coverage against calibration size (planted-risk world, 200 replications per n). The uniform band is the only column that certifies every k; the rest bind at one point or one grid. Rows where 10% of n falls below $k_{\min} = 100$ (here, $n = 500$) use $k = 100$ instead, i.e. 20% coverage for that row.

n	Thm 1	CP (fixed k)	Hoeffding	LTT-B10	LTT-B100	LTT-B1000
500	0.2529	0.0629	0.1224	0.1008	0.1323	0.1603
1,000	0.2466	0.0606	0.1224	0.0973	0.1282	0.1558
2,000	0.1432	0.0401	0.0865	0.0642	0.0841	0.1017
5,000	0.0761	0.0236	0.0547	0.0375	0.0488	0.0587
10,000	0.0487	0.0161	0.0387	0.0255	0.0330	0.0395
20,000	0.0322	0.0111	0.0274	0.0176	0.0227	0.0271
50,000	0.0192	0.0069	0.0173	0.0108	0.0139	0.0165

Table 7: Utility at the target level $\alpha = 0.10$, Tier C, $n_{\text{cal}} = 2{,}000$, pooled over 24 settings unless noted. "Feasible" is the fraction of replications with any certified operating point, not the fraction meeting $\alpha = 0.10$ specifically—Theorem 1's point always exists by construction ($k^\star = \arg\min_k U_k$) regardless of whether it meets the target; see Section 5.2's discussion of the resulting 0.554-at-51% point. The ARC-Easy row uses the same protocol restricted to that subset alone (Section 5.2); the coverage floor is Corollary 2.

Method	**Mean coverage**	**Mean held-out risk**	$\Pr[\text{risk} > \alpha]$	**Feasible**	**Cov. floor**
No certificate (pick on curve)	0.092	0.1145	0.740	0.11	—
Conformal risk control (in expectation)	0.091	0.1110	0.695	0.10	—
LTT fixed-sequence	0.074	0.0952	0.300	0.02	—
LTT-B100 grid	0.062	0.0872	0.000	0.00	—
Thm 3 population band	—	—	—	0.00	—
Thm 1 at $k^\star = \arg\min_k U_k$	0.510	0.4355	0.973	1.00	0.480
ARC-Easy subset only (N=5,027)	0.143	0.0827	—	1.00	—

Table 10: All 32 LLM Tier-A arms individually (model/group/score), $\delta = 0.05$, $k_{\min} = 100$. Every arm: 0/1000 violations.

Arm	**Reps**	**Rate**	**95% CI**	$p(> 0.05)$
phi-3.5-mini/all/verbal	1000	0.0000	[0.0000, 0.0037]	1
phi-3.5-mini/mmlu/entropy	1000	0.0000	[0.0000, 0.0037]	1
phi-3.5-mini/mmlu/hidcent	1000	0.0000	[0.0000, 0.0037]	1
phi-3.5-mini/mmlu/hidknn	1000	0.0000	[0.0000, 0.0037]	1
phi-3.5-mini/mmlu/maxprob	1000	0.0000	[0.0000, 0.0037]	1
phi-3.5-mini/mmlu/seqlp	1000	0.0000	[0.0000, 0.0037]	1
phi-3.5-mini/pool/entropy	1000	0.0000	[0.0000, 0.0037]	1
phi-3.5-mini/pool/hidcent	1000	0.0000	[0.0000, 0.0037]	1
phi-3.5-mini/pool/hidknn	1000	0.0000	[0.0000, 0.0037]	1
phi-3.5-mini/pool/maxprob	1000	0.0000	[0.0000, 0.0037]	1
phi-3.5-mini/pool/seqlp	1000	0.0000	[0.0000, 0.0037]	1
pythia-410m/mmlu/entropy	1000	0.0000	[0.0000, 0.0037]	1
pythia-410m/mmlu/hidcent	1000	0.0000	[0.0000, 0.0037]	1
pythia-410m/mmlu/hidknn	1000	0.0000	[0.0000, 0.0037]	1
pythia-410m/mmlu/maxprob	1000	0.0000	[0.0000, 0.0037]	1
pythia-410m/mmlu/seqlp	1000	0.0000	[0.0000, 0.0037]	1
pythia-410m/pool/entropy	1000	0.0000	[0.0000, 0.0037]	1
pythia-410m/pool/hidcent	1000	0.0000	[0.0000, 0.0037]	1
pythia-410m/pool/hidknn	1000	0.0000	[0.0000, 0.0037]	1
pythia-410m/pool/maxprob	1000	0.0000	[0.0000, 0.0037]	1
pythia-410m/pool/seqlp	1000	0.0000	[0.0000, 0.0037]	1
qwen2.5-1.5b/all/verbal	1000	0.0000	[0.0000, 0.0037]	1
qwen2.5-1.5b/mmlu/entropy	1000	0.0000	[0.0000, 0.0037]	1
qwen2.5-1.5b/mmlu/hidcent	1000	0.0000	[0.0000, 0.0037]	1
qwen2.5-1.5b/mmlu/hidknn	1000	0.0000	[0.0000, 0.0037]	1
qwen2.5-1.5b/mmlu/maxprob	1000	0.0000	[0.0000, 0.0037]	1
qwen2.5-1.5b/mmlu/seqlp	1000	0.0000	[0.0000, 0.0037]	1
qwen2.5-1.5b/pool/entropy	1000	0.0000	[0.0000, 0.0037]	1
qwen2.5-1.5b/pool/hidcent	1000	0.0000	[0.0000, 0.0037]	1
qwen2.5-1.5b/pool/hidknn	1000	0.0000	[0.0000, 0.0037]	1
qwen2.5-1.5b/pool/maxprob	1000	0.0000	[0.0000, 0.0037]	1
qwen2.5-1.5b/pool/seqlp	1000	0.0000	[0.0000, 0.0037]	1

Table 8: Exchangeability stress test. ResNet-18 calibrated on clean CIFAR-10 (90.8% test accuracy, mean $\eta = 0.115$) and deployed at CIFAR-10-C severity s; 300 replications each, $n_{\text{cal}} = 5{,}000$.

Severity	Violation rate	95% CI	mean-η gap	
0 (matched control)	0.003	$[0.000, 0.018]$	$+0.000$	within budget
1	1.000	$[0.988, 1.000]$	$+0.063$	broken
2	1.000	$[0.988, 1.000]$	$+0.123$	broken
3	1.000	$[0.988, 1.000]$	$+0.175$	broken
4	1.000	$[0.988, 1.000]$	$+0.245$	broken
5	1.000	$[0.988, 1.000]$	$+0.354$	broken

Table 9: Margin quantiles at the post-hoc operating point k_1, pooled over 24,000 replications. A negative margin means the realized risk sits below the certified bound (slack); Usage× is the violation rate divided by the $\delta = 0.05$ budget.

Certificate	Violation rate	Usage×	Median margin	P95 margin
Population band (Theorem 3)	0.0000	0.00	-0.5338	-0.0421
Transductive band (Theorem 1)	0.0000	0.00	-0.0999	-0.0602
Clopper–Pearson	0.1267	2.53	-0.0335	$+0.0178$

D.3 The betting bound before and after the patch

Table 11: Betting-bound (WSR) violation rate before and after the Lemma 1 patch. The calibration-side audit uses a single pre-registered $k_0 = 1{,}000$ with no selection and no population transfer, pooled over 3,000 replications across our three base models; the deployment rows repeat Table 1's protocol at $n_{\text{cal}} = 2{,}000$, pooled over 24,000 replications.

Setting	Raw	Patched	95% CI (patched)
Calibration-side audit (k_0 fixed, no transfer)	0.4197	0.0000	$[0.0000, 0.0012]$
Deployment, r_3 (fixed k_0, no selection)	0.3006	0.0478	$[0.0451, 0.0505]$
Deployment, r_1 (post-hoc argmin)	0.1145	0.0560	$[0.0532, 0.0590]$

The mechanism behind the raw rate: our implementation's bisection search begins at the sample mean, and if the underlying sequential test already rejects right there, it returns that starting point—the sample mean, zero margin—instead of continuing to search. This path fires in 32.2% of replications on our real, heterogeneous-risk calibration data, against roughly 1% on i.i.d. synthetic Bernoulli data of the same size; one plausible contributor is that our data is sorted by score before this test ever sees it, which builds in exactly the kind of systematic trend across the sequence that a genuinely i.i.d. draw does not have, though we have not isolated this as the specific cause. The gap between those two numbers is the gap between a rare, budgeted edge case and one that fires far more often than its own δ allows, and it's specifically heterogeneity in per-item risk—not selection, not population transfer—that separates them. We have not run this diagnostic against the officially released `confseq` package accompanying Waudby-Smith & Ramdas (2024); a reading of its public source shows a structurally different construction (a full scan over a discretized grid of candidate means, with no single starting point to exit early from), though we have not verified empirically that this difference is what matters.

D.4 Cluster-robust confidence intervals

Table 1's intervals are exact Clopper–Pearson on the pooled 24,000 replications, which is correct if those replications are independent. They aren't: within one of the 24 (model, group, score) settings, all 1,000 replications share the same η vector, the same score ordering, and the same underlying dataset, and only the loss realization gets redrawn. The unit of independent replication is really the setting—24 of them, not 24,000. We recompute intervals by resampling settings, not replications, with replacement, 10,000 bootstrap draws, under r_1.

The naive interval (Table 12) isn't wrong about the point estimate, only about how much to trust it, understating uncertainty by 2–8× depending on the certificate, worse for certificates whose failure rate itself varies more across settings. We report both throughout the released per-replication logs so either can be recomputed exactly.

D.5 Block-correlation sweep, full numbers

The LLM pool is blocked by its four constituent source datasets (ARC-Challenge, ARC-Easy, CommonsenseQA, OpenBookQA), verified against the released data-loading code's tag-to-block assignment rather than assumed from item counts. CIFAR-10 is blocked by its ten class labels, read directly from the dataset. Both give the same qualitative result on unrelated data modalities (full numbers in Table 13): item-level certification degrades with ρ, block-level certification doesn't move (Remark 3 spells out exactly what "block-level" certifies when, as here, the blocks aren't equal-sized).

Table 12: Naive pooled versus cluster-robust (setting-level bootstrap) 95% confidence intervals, r_1, $n_{\text{cal}} = 2{,}000$. CP's own per-setting rate ranges from 0.066 to 0.320 across the 24 settings—genuine heterogeneity, not sampling noise around a single shared rate, which is what the naive interval implicitly assumes.

Certificate	Naive pooled CI	Cluster-robust CI	Width ratio
Clopper–Pearson	$[0.1225, 0.1309]$	$[0.1062, 0.1512]$	$5.3\times$
Hoeffding	$[0.0193, 0.0229]$	$[0.0129, 0.0309]$	$4.9\times$
Empirical Bernstein	$[0.0008, 0.0017]$	$[0.0005, 0.0022]$	$1.9\times$
Betting / WSR (raw)	$[0.1117, 0.1198]$	$[0.0842, 0.1509]$	$8.2\times$

Table 13: Numeric backing for Figure 1(c)–(d): violation rate against induced within-block correlation ρ, 1,000 replications per cell, item- versus block-level certification on two independent block structures.

ρ	LLM, item	LLM, block	CIFAR, item	CIFAR, block
0.0	0.0000	0.0000	0.0000	0.0000
0.2	0.3890	0.0000	0.1000	0.0000
0.4	0.4300	0.0000	0.1740	0.0000
0.6	0.4600	0.0000	0.2720	0.0000
0.8	0.5200	0.0000	0.3020	0.0000

D.6 Necessity versus calibration size, full numbers

Table 14: Extending Section 5.3's single-n sweep across four calibration sizes, 500 replications per cell, same construction ($p = 0.15$, homogeneous, score independent of loss). The $n = 20{,}000$ column is an independent Monte Carlo draw from the main-body table's (different replication count, different seed); the two agree closely rather than exactly, as expected of two independent estimates of the same rate.

n	$c = 1.00$	$c = 0.75$	$c = 0.50$	$c = 0.25$
2,000	0.0000	0.0020	0.0460	0.3700
5,000	0.0000	0.0000	0.0600	0.4880
20,000	0.0000	0.0040	0.0980	0.6360
50,000	0.0000	0.0040	0.1200	0.6500

At $c = 0.50$ and $c = 0.25$ (Table 14), the violation rate rises monotonically with n across this $25\times$ range; at $c = 0.75$ it is non-monotonic within Monte Carlo noise (the rate is 0 at both $n = 2{,}000$ and $n = 5{,}000$ before rising), consistent with a $\ln \ln n$ mechanism that grows slowly rather than one that happens to bite at $n = 20{,}000$ specifically. This is evidence for the qualitative claim of Theorem 2, not a proof of a finite-sample rate; the discussion names the corresponding non-asymptotic result as the most valuable open theoretical direction.

E Datasets and models

E.1 Datasets

All four MCQ sources are standard, publicly released English-language benchmarks (Table 15) accessed through the `datasets` library (version in Appendix J); no new human data was collected. Options are scored by length-normalized log-likelihood (Gao et al., 2023), the standard scoring convention for these benchmarks. ARC and OpenBookQA test elementary/middle-school science knowledge; CommonsenseQA tests everyday commonsense reasoning; MMLU spans 57 subjects from elementary mathematics to professional law and medicine, used here as a held-out group distinct from the four-dataset pool specifically because its subject labels give the group structure Appendix F needs. TriviaQA supplies open-ended factual questions for Tier B, where genuine sampled decodes (rather than a simulated loss) are required. CIFAR-10-C applies 19 corruption types (noise, blur, weather, digital) at 5 severities to the CIFAR-10 test set, standard in the distribution-shift literature, and is used here purely as a controlled exchangeability-violation of known, increasing magnitude.

E.2 Models

The three language models (Table 16) span roughly an order of magnitude in parameter count and a 0.31–0.53 range in task accuracy—against a 22.8% chance floor on the pool's mix of 4-way and 5-way multiple choice—chosen to test whether the band's validity depends on model strength (Section 5.1 confirms it does not: the violation rate is exactly 0 for all three, and for the vision model). Pythia-410M is a base (non-instruct) model, scored by MCQ log-likelihood only; the two instruct models are additionally used for verbalized confidence and genuine-decode Tier B evaluation, which requires instruction-following to elicit either a confidence number or a sampled free-text answer. ResNet-18 is trained from scratch rather than pretrained, giving exact control over the train/test protocol needed for the exchangeability stress test and its own Tier-A matched control (Appendix D).

[3]The pool is capped at 24,000 items per replication draw (Appendix J); the four constituent datasets total 24,706 before this cap.

Table 15: Datasets used in this study, with the exact item counts loaded by the released pipeline. "Role" indicates where each dataset enters the experimental design (Section 4).

Dataset	Task	Items	Role
ARC-Easy (Clark et al., 2018)	4-way science MCQ	5,197	LLM pool constituent; utility subset (Section 5.2)
ARC-Challenge (Clark et al., 2018)	4-way science MCQ	2,590	LLM pool constituent
OpenBookQA (Mihaylov et al., 2018)	4-way science MCQ	5,957	LLM pool constituent
CommonsenseQA (Talmor et al., 2019)	5-way commonsense MCQ	10,962	LLM pool constituent
Pool total		24,706[3]	Tiers A/C, "pool" group
MMLU (Hendrycks et al., 2021)	4-way, 57 subjects	14,042	Separate group; group-conditional study (Appendix F)
TriviaQA (Joshi et al., 2017)	open-ended QA	4,000	Tier B (genuine decoding)
CIFAR-10 (Krizhevsky, 2009)	10-way image classification	50,000 / 10,000	Standard train/test split, ResNet-18 training
CIFAR-10-C (Hendrycks & Dietterich, 2019)	10-way, 19 corruptions $\times$ 5 severities	2,000/severity	Exchangeability stress test (Section 5.1)

Table 16: Models used in this study. Accuracy is on the held-out portion of the LLM pool at greedy decoding; all three language models are used off-the-shelf with no fine-tuning.

Model	Role	Params	Source / notes
Pythia-410M (Biderman et al., 2023)	Weak base LM	410M	EleutherAI suite; pool acc. 0.308
Qwen2.5-1.5B-Instruct (Yang et al., 2024)	Mid-strength instruct LM	1.5B	Pool acc. 0.509; also Tier B, verbalized confidence
Phi-3.5-mini-instruct (Abdin et al., 2024)	Strong instruct LM	3.8B	Pool acc. 0.527; also Tier B, verbalized confidence
ResNet-18 (He et al., 2016)	Vision classifier	11.2M	Trained from scratch on CIFAR-10, 20 epochs, batch 256; test acc. 0.908

F GROUP-CONDITIONAL VALIDITY

Theorem 1's guarantee is marginal: it controls the probability, over the randomness in the losses, that the band is violated somewhere. It says nothing directly about whether that probability is controlled *within* every pre-specified subgroup of the population simultaneously—a distinct question, and one that Jones et al. (2021) show can matter in selective classification specifically. The fix, if the effect is real, is standard: apply Theorem 1 separately to each of G groups at level δ/G and union-bound.

Proposition 1 (Group-conditional band)**.** *Let $B_1, \ldots, B_G$ be pre-specified, non-overlapping groups. Applying Theorem 1 to each group's own sorted sub-population at level δ/G gives a band simultaneously δ-valid within every group and at every within-group k.*

Proof. Immediate union bound over the G per-group applications of Theorem 1, each contributing failure probability at most δ/G. □

Proposition 1 isn't really the interesting question, because it's a tautology: a band separately calibrated and checked within a group is valid within that group by Theorem 1 applied verbatim, whether you call it group-conditional or not. The question worth measuring is different: when a single band is calibrated on the pooled population, as every method in this paper is, and a single deployment threshold gets chosen from it, can a specific subgroup's realized risk among the selected items exceed the pooled certificate even when the pooled realized risk doesn't?

We measured this on Phi-3.5-mini across all 57 MMLU subjects, comparing a pooled band built at the standard $\delta = 0.05$ against one built at $\delta/57$, 4,000 replications each, using the same r_1 selection rule as Section 5.4.

Table 17: Worst-subject held-out violation rate when a single pooled band is calibrated on all 57 MMLU subjects together and deployed at a post-hoc threshold, versus building the pooled band at the Bonferroni-corrected level. Pooled-overall violation is 0.0000 in both rows; the question is whether any single subject's realized risk among the selected items exceeds the pooled certificate even so.

Band level	Worst-subject violation	95% CI	Reps
$\delta = 0.05$ (marginal)	0.0018	$[0.0007, 0.0036]$	4,000
$\delta/57$ (Bonferroni)	0.0005	$[0.0001, 0.0018]$	4,000

The point estimate moves the right way (Table 17), but the two intervals overlap. We don't read this as a null result so much as an inconclusive one at this sample size: 57 subjects with a few hundred items each doesn't give much power to resolve an effect this size, and 4,000 replications isn't enough to separate something real-but-small from noise around it. We're reporting the numbers rather than a figure, because a figure would suggest a sharper contrast than the data actually support, and we'd rather understate this result than oversell it. A study purpose-built to resolve it—many more items per subject, a base rate chosen to make the effect larger—is the natural follow-up, and the discussion names it as such.

G SCORE-FAMILY CONTRASTS

Mean band width at 10% coverage, differenced across score families and tested by permutation with Holm correction (Holm, 1979) (10,000 permutations; the reported p is at the permutation floor). Positive means the first-named score is wider (less tight); each entry is computed as first minus second.

Contrast	Mean width difference	Holm-adjusted p
maxprob − entropy	-0.0001	$< 10^{-3}$
maxprob − seqlp	$+0.0131$	$< 10^{-3}$
maxprob − hidcent	$+0.0379$	$< 10^{-3}$
maxprob − hidknn	$+0.0387$	$< 10^{-3}$
seqlp − hidcent	$+0.0249$	$< 10^{-3}$
hidcent − hidknn	$+0.0008$	$< 10^{-3}$

Hidden-state scores order items better than output probabilities on these tasks, by roughly 0.038 in certified width. These six pairwise contrasts are also why we count hidden-state centroid distance and the k-NN score as two separate families rather than one throughout the paper (see Appendix C's note on the score-family count): they get their own dedicated comparison here, 'hidcent − hidknn', which wouldn't make sense if we didn't already treat them as distinct. Because Theorem 1 is score-agnostic, this whole table is a statement about representation quality that the certificate simply inherits, not about which certificate to trust.

H Additional figures

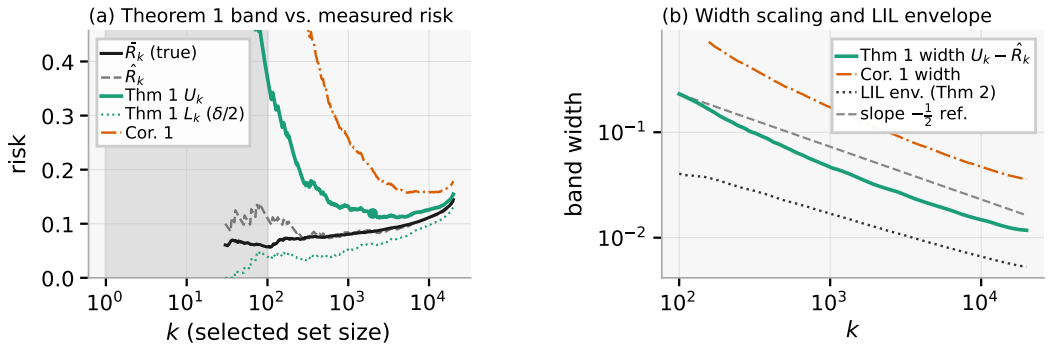


Figure 4: The band on one planted-risk replication ($n = 20{,}000$). **(a)** U_k dominates $\bar{R}_k$ at every prefix length; shading marks $k < k_{\min}$. **(b)** Width against k, with the iterated-logarithm envelope of Theorem 2 and a $-1/2$ slope reference. The band tracks the envelope rather than the parametric rate—the visual signature of uniformity.

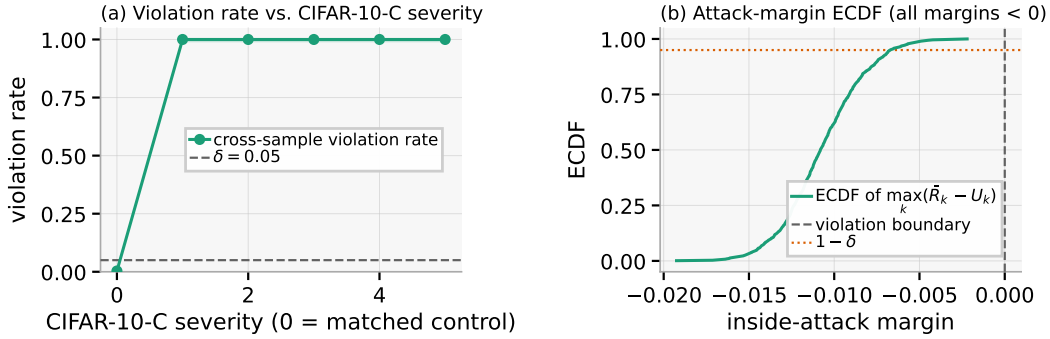


Figure 5: **(a)** Violation rate for a certificate calibrated on clean CIFAR-10 and deployed at CIFAR-10-C severity s. **(b)** Distribution of $\max_k(\bar{R}_k - U_k)$ in the planted-risk study; the entire mass sits left of zero, with median margin -0.011.

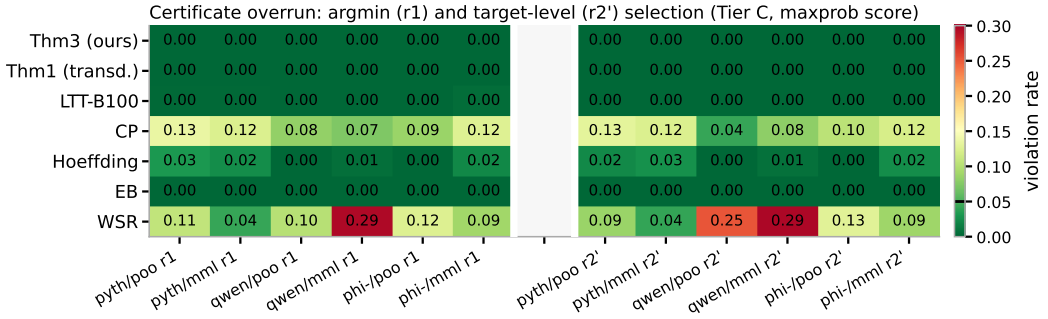


Figure 6: Held-out overrun rates (Tier C, $n_{\text{cal}} = 2{,}000$, 1,000 reps/cell, all 6 model×group pairs at maxprob shown individually). Left: selection at the empirical minimum (r_1). Right: selection at the prefix closest to target (r_2'). Clopper–Pearson and the betting bound break under either rule; LTT-B100 has no r_2' analogue, since it only certifies its own pre-registered grid (n/a). The pooled version of this same result is Table 1 in the main body; this figure gives the per-setting breakdown behind it.

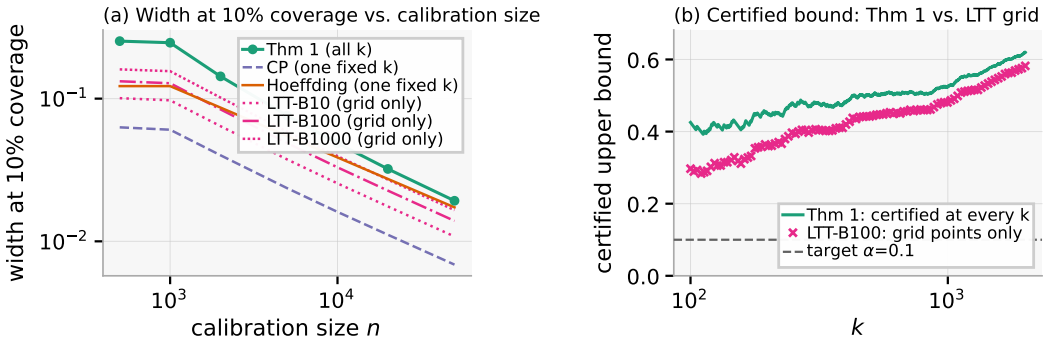


Figure 7: (a) Width at 10% coverage against calibration size for the uniform band and for certificates that bind at one point or one grid. **(b)** What each certifies: our band is defined at every k, LTT-B100 only at its grid points, and the fixed-sequence variant only on a prefix.

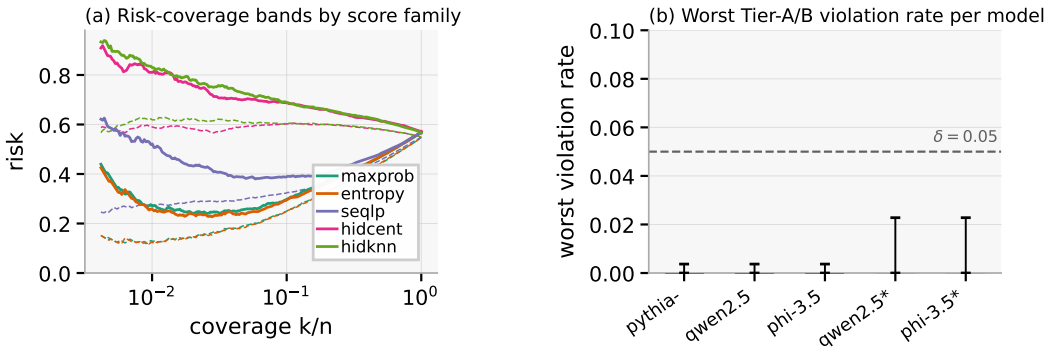


Figure 8: (a) Certified risk–coverage bands on Phi-3.5-mini across the score family. **(b)** Worst-arm violation rate per model; starred entries are Tier B, where the resolution floor is $1/32$.

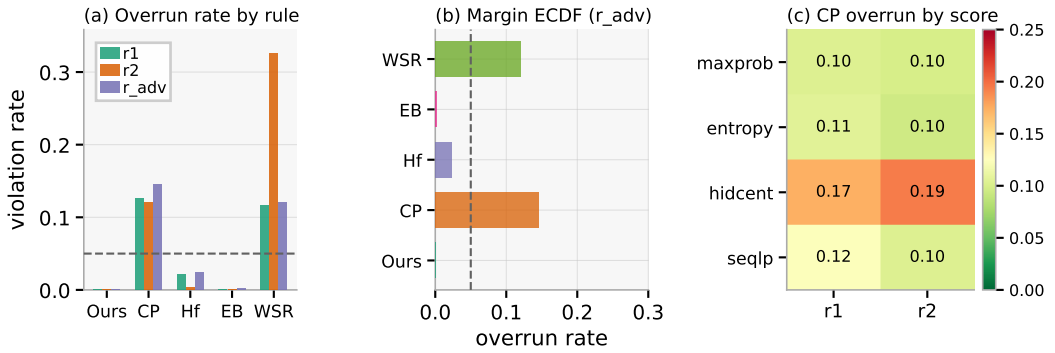


Figure 9: Overrun rate by certificate and selection rule at $n_{\text{cal}} = 2{,}000$. **(a)** Rates under r_1, r_2' and r_{adv}. **(b)** Rates under the adversarial rule alone, as an ECDF of the inside-attack margin. **(c)** Clopper–Pearson overrun broken out by confidence score under both r_1 and r_2'; the failure is not specific to any one score, and every cell is defined now that r_2' replaces the original rule.

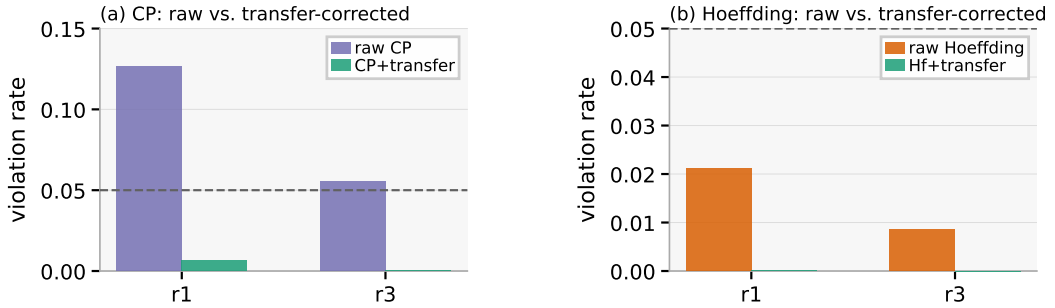


Figure 10: Isolating the population-transfer gap from selection, as described in Section 5.4: raw versus transfer-corrected violation rate for Clopper–Pearson **(a)** and Hoeffding **(b)**, under r_1 and r_3. The correction removes essentially all of both certificates' failure at this scale, isolating population transfer as the dominant term rather than selection specifically.

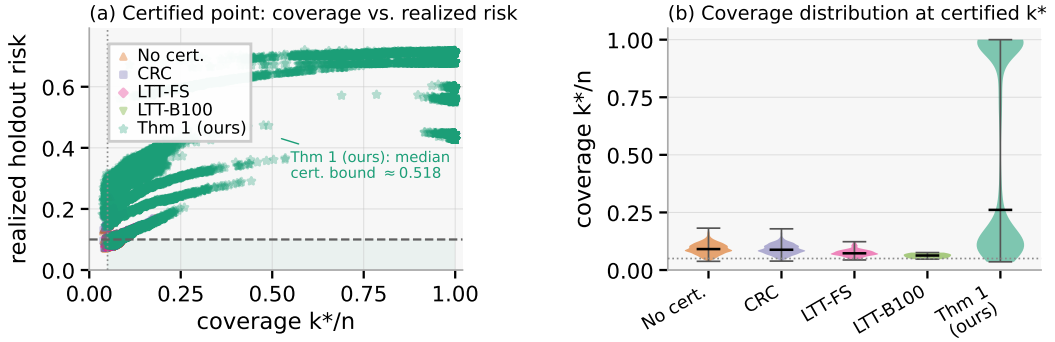


Figure 11: Operating points in coverage–risk space at $n_{\text{cal}} = 2{,}000$. The shaded band is the region a practitioner wants (coverage above 5%, risk at or below $\alpha = 0.10$). Methods without a high-probability guarantee scatter above the target line; Theorem 1's transductive operating point (starred) sits at high coverage and high risk, the certified-but-unappealing point Section 5.2 discusses; the population band of Theorem 3 produces no point at all in this regime, which is Corollary 1 binding rather than a numerical failure.

I FULL SCALING-EXPERIMENT NUMBERS

At $n_{\text{cal}} = 20{,}000$ on Phi-3.5-mini, 500 replications, every baseline's rate at the pre-registered k_0 rose sharply: Clopper–Pearson to 0.244, Hoeffding to 0.144, empirical Bernstein to 0.118, betting to 0.536, LTT-B100 to 0.092, against 0.008 for ours. The pool leaves only $4{,}000$ held-out items after this split, about $2{,}000$ of them above threshold, giving a held-out standard error of 0.0107—the same order as the certificate widths under test (0.0080 Clopper–Pearson, 0.0122 Hoeffding, 0.0139 empirical Bernstein). An exactly tight certificate would be exceeded by measurement noise alone at rates 0.229, 0.127 and 0.098 respectively, within a couple of points of the observed 0.244, 0.144 and 0.118. The experiment cannot separate an invalid certificate from a valid one measured with an undersized holdout, and we draw no conclusion from it either way.

J REPRODUCIBILITY

Every arm is checkpointed at the replication level and every reported number is recomputable from released per-replication CSVs. Seeds derive deterministically from a single master seed via a counter-based generator, so a replication is identified by (arm id, replication index) and is reproducible in isolation. The run manifest records package versions (`numpy` 2.0.2, `scipy` 1.16.3, `torch` 2.10.0+cu128, `transformers` 4.57.6), the hardware (a single Tesla T4), and wall-clock time (6.3 GPU-hours for the complete study including the second-wave additions described in Appendix C, of which language-model scoring accounts for approximately 5.9; every result from that wave is CPU-only and the full extended suite runs in under fifteen minutes on a standard notebook CPU). All baseline certificates — Clopper–Pearson, Hoeffding, empirical Bernstein, the betting bound, and LTT — are implemented from scratch in this codebase rather than calling external statistical packages; Appendix D.3 discusses a consequence of this choice for the betting bound specifically. Three self-checks run inside the notebook and abort the study on failure: an RNG-stream reproducibility check, a forward-pass determinism check on the loaded model, and a full-pipeline determinism check that runs the same seed twice and compares arrays elementwise. The assumption audits reported in Section 4 are executed at run time, not asserted: the cross-replicate residual correlation is -0.0080 and the adjacent-item residual correlation is $+0.0041$ against a threshold of $4/\sqrt{n} = 0.0283$. Every result added in the second wave was independently reproduced end to end on a second machine from the released checkpoints before being written into the paper.

J.1 COMPLETE HYPERPARAMETER AND REPRODUCIBILITY REFERENCE

Table 18 collects every parameter governing the study in one place: statistical protocol, per-arm replication counts, model and dataset sizes, score-family settings, and hardware/software. Every value is taken directly from the released run manifest or the notebook's own configuration dictionary, not retyped from memory. Where a parameter varies across an ablation (marked $\star$), the swept set is given; the value used in every other reported run is the one listed.

Table 18: Complete hyperparameter and reproducibility reference. All values are taken verbatim from the released run manifest and notebook configuration; none are reconstructed from memory. $\star$SWEEP: the parameter is varied in the named sweep or ablation; the listed value is used in every other reported run.

Parameter	Value	Notes
Statistical protocol		
Confidence level δ	0.05	fixed throughout; the DKW/population constant also uses this δ
Minimum prefix $k_{\min}$	100	band and all verdicts evaluated only for $k \geq k_{\min}$
Target risk level α	0.10	feasibility (Corollary 1) and deployment protocol (Section 5.4)
Geometric base b	2	λ_j-menu stitching, Eq. (3); $\star$ABL: also tested at $b \in \{e, 1.5\}$
δ-allocation across menu	polynomial $(6/\pi^2(j{+}1)^2)$	$\star$ABL: also tested with geometric allocation
Score tie-breaking	i.i.d. $\mathrm{Unif}(0,1)$ jitter, 10^{-9} scale	applied once per replication before sorting
Language models		
Pythia-410M	410M params	Biderman et al. (2023); base (non-instruct)
Qwen2.5-1.5B-Instruct	1.5B params	Yang et al. (2024)
Phi-3.5-mini-instruct	3.8B params	Abdin et al. (2024)

continued on next page

Table 18 – continued from previous page

Parameter	Value	Notes
Vision model	ResNet-18, 11.2M params	He et al. (2016); trained from scratch
CIFAR training epochs	20	batch size 256
CIFAR-10-C items per severity	2,000	19 corruption types $\times$ 5 severities (Hendrycks & Dietterich, 2019)
Datasets and pool construction		
ARC-Easy / ARC-Challenge	5,197 / 2,590 items	Clark et al. (2018); pool constituents
OpenBookQA / CommonsenseQA	5,957 / 10,962 items	Mihaylov et al. (2018); Talmor et al. (2019); pool constituents
Pool cap	24,000 items	drawn from 24,706 available across the four sources
MMLU (separate group)	14,042 items, 57 subjects	Hendrycks et al. (2021); group-conditional study
TriviaQA subset	4,000 questions	Joshi et al. (2017); Tier B
Replication counts, by arm		
Synthetic validity (V1/E1)	1,000 reps	$n = 20{,}000$ planted-risk world
Necessity ablations	300 reps	sub-LIL width fractions
Real-activation Tier A (E2)	1,000 reps/arm	32 LLM arms + 1 CIFAR matched-control arm (300 reps)
Genuine-decode Tier B (E3)	32 samples/question, 1,000 bootstrap reps	10 (model, score) arms
Deployment (E4)	1,000 reps/setting	24 settings at $n_{\text{cal}} = 2{,}000$
Exchangeability stress (E6)	300 reps/severity	$n_{\text{cal}} = 5{,}000$
Block-correlation sweep	1,000 reps/cell	5 ρ values $\times$ 2 block structures $\times$ 2 levels (item/block)
Necessity-vs-n sweep	500 reps/cell	4 sizes $\times$ 4 width fractions
Group-conditional (MMLU)	4,000 reps	57 subjects, marginal vs. Bonferroni-corrected
Band-width n-sweep	200 reps/point	$n \in \{500, \ldots, 50{,}000\}$, 7 points
Deployment protocol (E4) settings		
Primary calibration size n_{cal}	2,000	$\star$SWEEP: also $\{1{,}000, 5{,}000, 20{,}000\}$
Fixed threshold k_0 (rule r_3)	$n_{\text{cal}}/2$	pre-registered, no selection
Score families tested	maxprob, entropy, hidcent, seqlp	4 of the 7 total score families
LTT grid sizes	$\{10, 100, 1{,}000\}$	Bonferroni-corrected grids, own-point comparison
Held-out set size	$\geq$ 3,000 items	remainder of the pool after calibration draw
Score-family parameters		
Hidden-state k-NN dimension	256	PCA-reduced hidden state
k-NN reference set / k	8,000 items / k=10	held out from scoring set
Verbalized-confidence subset	8,000 prompts, both instruct models	max 4 new tokens
Tier B decode samples	32/question, max 24 new tokens	TriviaQA, T=1
Seeds and hardware		
Master seed	20260727	fixed for the entire study
RNG scheme	counter-based, keyed (arm id, rep index)	every replication reproducible in isolation
RNG scope	NumPy, PyTorch CPU and CUDA	`cudnn.deterministic=True`
Hardware	single NVIDIA Tesla T4	16GB VRAM
Python	3.12.13	
Key packages	`numpy` 2.0.2, `scipy` 1.16.3, `torch` 2.10.0+cu128 `transformers` 4.57.6, `datasets` 5.0.0, `torchvision` 0.25.0+cu128	

J.2 REPRODUCIBILITY CROSSWALK

Table 19 traces every figure and table in this paper to the exact notebook cell and checkpoint file(s) that produced it, so that any reported number can be located and re-run without searching.

Table 19: Crosswalk from every reported figure and table to its exact source cell and checkpoint file(s).

Paper item	Notebook cell id	Source checkpoint(s)	Location
Table 1	`s7-stats2 / s7-verdicts2`	`e4_*.npz`	main body
Figure 1	`fig-necessity-blockrobust-combined`	`necessity_and_blockrobust_combined.npz`	main body
Table 2	`s6-feas`	`feas_table.npz`	Appendix B
Figure 2	`fig-feas-phase`	`feas_phase.npz`	Appendix B.1
Table 3	`s5-synth`	`v1_synthetic.npz`	Appendix D
CIFAR Tier-A row (Appendix D)	`s8-e6`	`e6_clean_tierA.npz`	Appendix D
Table 9	`s7-verdicts2`	`e4_*.npz`, `e4t2u_*.npz`	Appendix D.1
Figure 3	`s7-verdicts2 (ext.)`	`e4_*.npz`, `e4t2u_*.npz`	Appendix D.1
Table 10	`s7-verdicts2 (ext.)`	`e2_*.npz`	Appendix D.2
Table 11	`s7-wsr-audit / s7-verdicts2`	`wsr_audit_*.npz`	Appendix D.3
Table 12	`s7-cluster-boot`	`e4_*.npz`	Appendix D.4
Table 13	`s9-blockcorr full`	`blockcorr_*.npz`	Appendix D.5
Table 14	`s9-necscale-ful`	`necscale_*.npz`	Appendix D.6
Table 15	`s0-datasets`	`dataset manifest.json`	Appendix E
Table 16	`s0-models`	`model manifest.json`	Appendix E
Table 17	`s11-group-mmlu`	`mmlu_group_*.npz`	Appendix F
Figure 8	`fig-realdata`	`e2_*.npz`	Appendix H
Figure 4	`5348d9b0-..d828`	`curves_store.npz`	Appendix H
Figure 7	`fig-price-unif`	`bandwidth nsweep.npz`	Appendix H
Figure 5	`fig-robust stres`	`e6_*.npz`, `v1_margins.npz`	Appendix H
Figure 6	`7ec39717-...fc2`	`e4_*.npz`, `e4ext_*.npz`	Appendix H
Figure 9	`fig-fragility`	`e4_*.npz`	Appendix H
Figure 10	`s7-transfer-iso`	`e4_transfer_*.npz`	Appendix H
Figure 11	`fig-utility`	`e4_*.npz`	Appendix H
Table 18	`s10-pack2`	`run_manifest.json`	Appendix J.1

K Extended related work

Section 6 of the main text is deliberately compressed to fit the page budget. This appendix gives the fuller picture, including one piece of independent, concurrent evidence that was not available when the main text was drafted.

Independent, concurrent corroboration. Zhou & Wang (2026) audit the same family of certificates — Hoeffding, Clopper–Pearson, and the WSR betting bound — for selective risk control in a completely different application domain: machine anomalous-sound detection and AI-generated-image forensics. Their protocol differs from ours in an important way worth stating precisely, since a careless reading would over-claim the connection: their certified rules apply a Bonferroni correction δ/n over every candidate threshold in the search, making them closer in spirit to our grid-corrected LTT baselines than to the uncorrected, single-point Clopper–Pearson/Hoeffding/WSR we study. Read this way, their two central findings align with ours cleanly rather than loosely. First, their uncertified *Naive* rule — empirical thresholding with no correction and no certificate at all — exceeds its declared budget in 49–73% of synthetic trials and up to 68% of real-data splits; our own uncertified baseline (Appendix D, "No certificate") overruns in 74% of replications, an independent, cross-domain convergence on the same qualitative phenomenon. Second, their Bonferroni-corrected certified rules stay within budget under exchangeable splits but break — 9–30% overrun — under a structural, grouped violation of exchangeability (calibrating on some categories, deploying on disjoint unseen ones), which they trace to a specific hard subgroup dominating the failure rather than to the bounds themselves. This is a different axis of exchangeability violation than our own CIFAR-10-C severity sweep (Section 5.1) — theirs is categorical (unseen group), ours is a continuous corruption magnitude — and the two together suggest the failure mode is general across how exchangeability breaks, not an artifact of either sweep's specific construction. Neither their audit nor ours tests what we consider the central failure mode of this paper: a fixed-k certificate evaluated at a threshold chosen *after* seeing the calibration curve, with no Bonferroni correction for that choice. That gap is, as far as we are aware, still open outside this work.

Distribution-free risk control, in more depth. Beyond the comparison in the main text, the risk-controlling prediction set framework of Bates et al. (2021) and Learn-then-Test (Angelopoulos et al., 2025) both certify a pre-registered, finite candidate set via a multiplicity correction; conformal risk control (Angelopoulos et al., 2024) relaxes the binary loss assumption to bounded, monotone losses at the cost of an expectation-level (rather than high-probability) guarantee, which is why we report it in Section 5.2's utility comparison as a distinct category rather than a peer of the high-probability methods. Adaptive and group-conditional conformal variants (Romano et al., 2020; Barber et al., 2021) relax the requirement that a single threshold serve every subgroup, at the cost of needing group labels at calibration time — the same trade-off Appendix F's Proposition 1 makes explicit and measures directly rather than assumes.

Anytime-valid inference, in more depth. The confidence-sequence literature (Darling & Robbins, 1967; Robbins, 1970) predates and anticipates much of the machinery we use; the non-asymptotic, finite-time forms we build on directly are due to Howard et al. (2020; 2021), whose line-crossing and stitching techniques give the λ_j-menu construction behind Theorem 1. The betting/e-value perspective (Waudby-Smith & Ramdas, 2024; Ramdas et al., 2023) reframes the same guarantees as wealth processes; Jamieson et al. (2014) is the clearest statement, in the best-arm-identification literature, of the iterated-logarithm rate as a floor rather than an artifact, which is the statement Theorem 2 transplants into the risk–coverage setting.

Selective prediction, in more depth. Beyond Chow's original reject option (Chow, 1970) and its foundational treatment (El-Yaniv & Wiener, 2010), selective classification with learned rejectors (Geifman & El-Yaniv, 2017) and its known miscalibration under distribution shift (Geifman et al., 2019) motivate why a distribution-free guarantee, not a learned confidence estimate, is the object worth certifying. Jones et al. (2021) is the closest prior identification of the specific multi-group failure mode Section 5.5 constructs directly rather than surveys.

Selection-based and stability-based risk control. A related line attacks data-driven selection or non-monotonic risk directly rather than via time-uniform concentration: Jin & Candès (2023) control false discovery rates under data-driven candidate selection via conformal p-values, and Angelopoulos (2026) extend conformal risk control to non-monotonic losses via leave-one-out algorithmic stability. Both trade the time-uniform, threshold-free guarantee we study for a different one (FDR control, or risk control for a fixed non-monotonic loss), rather than directly competing on the post-hoc prefix-selection problem this paper addresses.

Uncertainty estimation in language models, in more depth. Semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) and black-box sampling estimators (Lin et al., 2024) extend self-consistency-style scoring beyond exact-match agreement; verbalized confidence (Kadavath et al., 2022; Tian et al., 2023) elicits a number directly from the model, which we use as one of our score families precisely

because Theorem 1 requires no assumption about how that number was produced. Hidden-state probing (Azaria & Mitchell, 2023; Burns et al., 2023) is the origin of the centroid and k-NN scores that Appendix G finds empirically tightest on our pool — a finding about representation quality that the certificate inherits rather than assumes.