LEARN THE SHAPE, CALIBRATE THE SCALE: SIMULATOR-TRAINED CONFORMAL PREDICTION UNDER SIM-TO-REAL SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction provides finite-sample coverage guarantees, but adaptive intervals are difficult to deploy when labeled target data are scarce, even when a simulator can provide abundant synthetic data. We propose SUIT, a two-stage framework that learns the input-dependent uncertainty shape from the simulator and uses a small number of real samples to calibrate a scalar scale correction. This decomposition separates transferable uncertainty structure from the sim-to-real gap: zero-shot prediction is possible when the simulator is faithful, scalar calibration removes arbitrary global scale shifts, and stratification controls the residual spatial heterogeneity of the scale ratio. A conservative, computable audit of zero-shot degradation provides a stratum-wise lower bound on coverage from a small number of real audit points and flags inconclusive strata for recalibration. Experiments on distinct-value estimation, species-richness estimation, simulation-based inference, and synthetic regression show near-nominal zero-shot coverage under faithful simulation and coverage recovery with only 100 real samples under a 10× scale mismatch, at substantially narrower intervals than global conformal baselines.

1 INTRODUCTION

Conformal prediction Vovk et al. (2005); Papadopoulos et al. (2002); Lei et al. (2018); Angelopoulos & Bates (2023) upgrades any point predictor to prediction intervals with distribution-free marginal coverage. For heteroscedastic problems, two constraints stand between theory and deployment. The first is *adaptivity*: width that tracks the local noise level requires a trained adaptive component Romano et al. (2019); Guan (2023); Kiyani et al. (2024), fitted on labeled target-domain data and, for quantile methods, refitted at each coverage level α. The second constraint is *calibration*: the conformal step itself consumes a held-out labeled sample from the deployment distribution, so every new domain, and for the methods above every new α, demands its own calibration budget of n_{cal} labeled real points. When $n_{\text{cal}} = 0$, neither constraint can be met: there is nothing from which to train the adaptive component and nothing with which to calibrate.

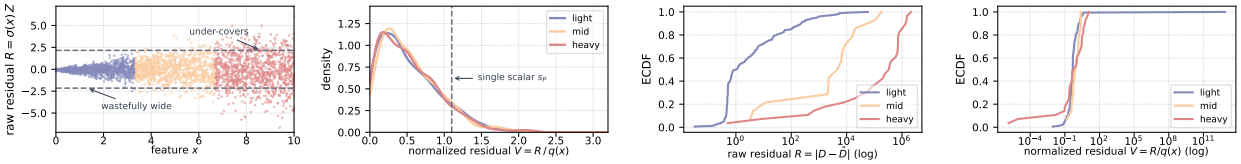
Yet in many domains the data-generating process is itself available as a stochastic simulator, from ecological species-accumulation models Good & Toulmin (1956); Chao (1984) to simulation-based inference Tejero-Cantero et al. (2020); Lueckmann et al. (2021). A concrete instance anchors the rest of the paper: estimating the number of distinct values in a database column from a small sample comes with an urn (occupancy) model of random sampling as an exact, tunable simulator, while each real label costs a full scan of the column (Appendix D). Simulators supply unlimited labeled synthetic data but do not dissolve the constraints: calibrating *inside* the simulator transfers any sim-to-real gap into coverage error, and existing simulator-based conformal pipelines Cabezas et al. (2025); Patel et al. (2023) inherit simulator fidelity silently. This paper asks whether a simulator can be exploited *safely*: what to learn inside it, what to estimate from real data, and whether the residual damage can be audited.

Our answer rests on a decomposition: under a multiplicative heteroscedastic error model, predictive uncertainty factors as *shape* $\times$ *scalar*, and the two factors have sharply different statistical status (Figure 1). The factorization is not only a modeling convenience: on real data the raw residual is wildly heteroscedastic across noise regimes while the normalized residual is approximately shape-stable (Figure 1b,c). The shape, a conditional scale function $q(x)$, is high-dimensional but *amortizable from the simulator*: one training run serves all coverage levels α and all deployment domains. The scalar, one threshold multiplying the shape, is estimated from real data and absorbs the sim-to-real gap.

This paper develops SUIT (Simulator-trained Uncertainty Intervals via Two-stage calibration) around this decomposition. Our contributions:

(i) **The SUIT framework: learn the shape in the simulator, calibrate a scalar on reality** (Section 3). A two-stage framework splitting predictive uncertainty into a conditional scale function $q(x)$, which is trained once on

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.



(a) raw: no single scalar works (b) normalized: one scalar everywhere (c) real data: wildly heteroscedastic (d) real data: shape-stable after $q(x)$

Figure 1: Validation of shape–scale decomposition. (a, b) Synthetic data: Signed residuals $R = \sigma(x)Z$ show stratum-dependent spread under a global band (a), but collapse to a shared density $V = R/q(x)$ after scaling by $q(x)$, enabling a single s_P (b). (c, d) Real NDV data (215 columns): Stratified by tail index, raw errors $R = |D - \hat{D}|$ span five orders of magnitude (c), whereas normalized residual $V = R/q(x)$ collapse to a common bulk (d).

simulator samples, serving every coverage level α and deployment domain; and a scalar threshold carrying the entire sim-to-real correction, estimated from a dimension-free number of real points or skipped in zero-shot mode, with a conservative per-stratum audit flagging which strata require real-label calibration.

(ii) **Why the sim-to-real gap reduces to a scalar** (Section 4). Coverage loss is governed by the *dispersion* of the log scale-ratio $\lambda(x)$ between real and simulated noise scales; its *level* is removed exactly by scalar calibration (Theorems 1–2), so $O(\delta^{-2})$ real points restore pointwise near-conditional coverage under scale misspecification of arbitrary magnitude (Corollary 2), and coverage is exact at $n_{\text{cal}} = 0$ under pure covariate shift, even with disjoint supports, provided the learned scale extrapolates structurally to the real support (Corollary 1); global raw-score thresholds and weighted conformal prediction lack both properties (Propositions 1–2).

(iii) **A conservative deployment audit for zero-shot prediction** (Theorems 4–8). The log normalized score separates the simulator-shaped noise from the residual sim-to-real scale variation, enabling a computable, stratum-wise lower bound on zero-shot coverage from a small number of real audit points. The certificate is informative when the simulator is sufficiently faithful and becomes vacuous under severe mismatch, thereby providing a conservative trigger for selective recalibration. A KS-based audit additionally flags potential shape mismatch.

(iv) **Experiments on four data families** (Section 5). On distinct-value estimation over seven real databases, species-richness surveys, two simulation-based-inference benchmarks, and synthetic regression with controlled sim-to-real gaps, the zero-shot mode attains nominal coverage where the simulator is faithful, 100 real points restore coverage under a ten-fold scale misspecification, and the audit tracks realized degradation, while one scale model serves all coverage levels as quantile baselines retrain per level.

2 Problem Setup and Related Work

2.1 Problem Setup

Let $\mathbb{P}$ and $\mathbb{Q}$ denote the joint distributions of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} \subseteq \mathbb{R}$, under a *simulator* and a *real* environment. A predictor $h : \mathcal{X} \to \mathcal{Y}$ is fixed, and we make no assumption on its accuracy. We quantify uncertainty through the absolute residual $R := |Y - h(X)|$ and focus on the heteroscedastic regime, where its magnitude varies substantially across $x \in \mathcal{X}$ and therefore requires *adaptive* interval widths. We assume query access to a stochastic simulator that can generate arbitrarily many i.i.d. labeled samples (X, Y) together with predictions $h(X)$ under $(X, Y) \sim \mathbb{P}$. The real environment may differ from the simulator in both covariates and residual scale; Section 4 formalizes this sim-to-real gap through a log scale-ratio $\lambda(x)$. Given a miscoverage level $\alpha \in (0, 1)$, we seek an interval $C_\alpha(x) = \big[\, h(x) - s \cdot q(x),\ h(x) + s \cdot q(x) \,\big]$, where $q : \mathcal{X} \to \mathbb{R} * +$ is an adaptive conditional scale and $s > 0$ is a scalar threshold, such that the real-environment marginal coverage $\text{Cov}_{\text{real}}(C_\alpha) := \mathbb{Q}(Y \in C_\alpha(X))$ is close to $1 - \alpha$. Exact conditional coverage is impossible distribution-free Vovk et al. (2005); Lei et al. (2018); we therefore target marginal coverage, with conditional coverage as a refinement. Two requirements distinguish our setting from standard conformal prediction: the requested α may vary at deployment time, so learned components should not be retrained per α, and they should transfer across deployment domains. We consider zero-shot ($n_{\text{cal}} = 0$) and few-shot (n_{cal} of order 10–10^2) real calibration.

Standard conformal prediction provides marginal coverage through a global threshold, but is not adaptive under heteroscedasticity. Existing adaptive methods require labeled target data, while shift-robust alternatives rely on discrepancies that may be difficult to estimate or assess at deployment. We refer to these two limitations as **(P1)** *adaptivity costs target-domain training* and **(P2)** *shift-robust calibration is either label-hungry or unverifiable*.

2.2 Related Work

Target-trained adaptive CP. Adaptive conformal methods learn input-dependent uncertainty from labeled target data: CQR Romano et al. (2019) conformalizes two conditional quantile regressors (refitting per α), LCP Guan (2023)

Table 1: Comparison of adaptive interval methods.

Family	Method	Adaptive shape	Uses simulator	Real labels	Scale shift	Zero-shot	Cond. ctrl.
Conformal (fixed width)	SCP Papadopoulos et al. (2002); Lei et al. (2018)	–	–	✓	✓	–	–
	SSCP Papadopoulos et al. (2002); Lei et al. (2018)	coarse	–	✓	✓	–	coarse
Target-trained (needs labels)	LCP Guan (2023)	✓	–	✓	✓	–	✓
	CQRD / CQRF Romano et al. (2019)	✓	–	✓	✓	–	✓
Simulator-/ shift-based (gap uncontrolled)	WCP Tibshirani et al. (2019)	–	–	✓	partial	–	–
	CQRS Romano et al. (2019)	✓	✓	✓	✓	–	✓
	CQSS Cabezas et al. (2025)	✓	✓	–	–	✓	limited
–	SUIT (ours)	✓	✓	0 or few	✓	✓	✓

Legend: *Adaptive shape*: interval width tracks the local noise scale (coarse = piecewise-constant across a few strata); *Real labels*: target-domain labeled data required (0 or few = zero-shot / $O(10^2)$ for scalar recalibration); *Scale shift*: coverage maintained when the real noise scale departs from the simulator's (partial = restricted shift model, e.g., covariate shift only; trivially satisfied by simulator-free rows); *Zero-shot*: defined and informative at $n_{\mathrm{cal}} = 0$; *Cond. ctrl.*: per-regime coverage control on real data (coarse = per-stratum only; limited = partial or unverified).

reweights scores by a learned localizer. They provide adaptivity but incur target-domain training and calibration costs, making them unsuitable for zero-shot deployment and challenging in the few-shot regime.

Simulator- and shift-based CP. Weighted conformal reweights calibration scores by the likelihood ratio $w = d\mathbb{Q}^X / d\mathbb{P}^X$ Tibshirani et al. (2019); Foygel Barber et al. (2023), with follow-ups on estimation, drift, label shift, and risk control Yang et al. (2024); Qiu et al. (2023); Cauchois et al. (2024); Gibbs & Candès (2021); Podkopaev & Ramdas (2021); Angelopoulos et al. (2024); Farinhas et al. (2024); Xu et al. (2025); Aolaritei et al. (2025); but w may be unbounded or undefined under severe shift, and the weighted threshold remains a non-adaptive scalar. Simulation-based inference supplies simulators and benchmarks Tejero-Cantero et al. (2020); Lueckmann et al. (2021), and conformal calibration has entered this pipeline via CANVI Patel et al. (2023) and the locally adaptive CP4SBI Cabezas et al. (2025), with adjacent ideas in high-dimensional statistics, robot control, and traffic prediction LeRoy & Zhao (2021); Rahaman et al. (2026); Ma & Yang (2025); these approaches remove target-domain training but either retain a global threshold or assume the simulator is representative of the target distribution. CQRS Romano et al. (2019) is closest to SUIT, as it learns adaptive shape in the simulator and calibrates on real data; SUIT differs by explicitly decomposing the residual sim-to-real gap into an absorbable scale level and within-stratum dispersion, yielding the theoretical coverage characterization of Section 4 and the deployment audit of Theorem 4.

Other approaches. Meta-conformal methods learn calibration rules across auxiliary tasks Fisch et al. (2021); Park et al. (2022); source-free and pseudo-label calibration replace missing target labels with model predictions Angelman et al. (2025); Siahkali et al. (2026); Flechsig & Pilz (2025); and unseen-species or occupancy models supply a complementary application literature Good & Toulmin (1956); Chao (1984); Sesia et al. (2023). None of these studies simulator-trained adaptive intervals with explicit sim-to-real degradation control, and the data assumptions of the first two families do not instantiate the simulator-plus-real-calibration protocol studied here, so they are omitted from the quantitative comparison.

Table 1 orders this design space as a progression: each row removes one limitation of the rows above at the price of a new assumption, and each row is instantiated by one baseline of Section 5.1. The conformal controls are valid at any calibration budget but rigid; the target-trained rows buy a learned shape with target labels; the simulator- and shift-based rows remove the target-training bill but leave the residual sim-to-real gap uncontrolled. SUIT (Section 3) keeps CQRS's division of labor—shape amortized in the simulator, scalars calibrated on real data—and fills the two cells no row above can: it is defined and informative at $n_{\mathrm{cal}} = 0$, and its residual degradation is certifiable from simulator quantities (Theorem 4).

3 SUIT: Learn the Shape, Calibrate the Scale

SUIT builds intervals $C(x; s) = [\, h(x) \pm s\, q(x) \,]$: all design freedom resides in a *shape* q (an adaptive conditional scale) and a *scalar* s that yields a $(1 - \alpha)$ interval. The shape is high-dimensional but learnable on the simulator, where unlimited samples let one run amortize over all α and domains; the scalar carries the sim-to-real gap and is a one-dimensional quantile computation.

3.1 Core decomposition

Under the multiplicative error model of Section 2, the real-environment residual factors as

$$R_{\mathrm{real}}(x) \;=\; \underbrace{q(x)}_{\text{simulator-learned scale}} \cdot \underbrace{e^{\lambda(x)}}_{\text{sim-to-real gap}} \cdot \underbrace{Z/c}_{\text{standardized shape}}, \qquad \lambda(x) \;:=\; \log \frac{\sigma_{\mathrm{real}}(x)}{\sigma_{\mathrm{sim}}(x)}, \tag{1}$$

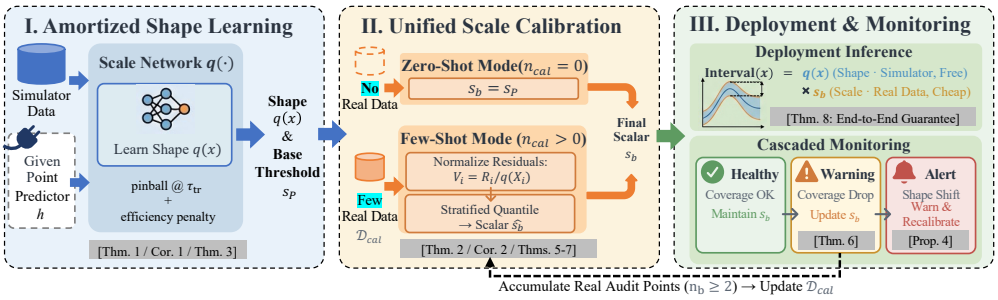


Figure 2: Overview of the SUIT framework.

where $q(x) = c\,\sigma_{\text{sim}}(x)$ is the conditional scale learned on the simulator and $Z \geq 0$, $Z \perp X$, is the standardized residual shape. We learn the x-dependent scale $q(x)$ from unlimited simulator data and use a small amount of real data to estimate the remaining scalar scale correction. Normalizing by the learned shape,

$$V = R/q(X) = e^{\lambda(X)} \cdot Z/c, \tag{2}$$

collapses the heteroscedastic structure: V is homogeneous up to the scalar field $\lambda(x)$ alone, which is exactly what a per-stratum scalar can absorb. The deployed interval is

$$C(x) = \big[\, h(x) - s_b\, q(x),\ h(x) + s_b\, q(x) \,\big], \tag{3}$$

with a single scalar s_b per stratum b: the shape q makes widths adaptive across $\mathcal{X}$, and the scalar s_b restores coverage in the real environment.

3.2 Three-step procedure

Figure 2 summarizes the pipeline (pseudocode in Algorithm 1, Appendix E).

Stage I — Shape learning. Train q once on simulator residuals $R = |Y - h(X)|$ with a pinball loss at a fixed training level plus an efficiency penalty that restores the coverage–width trade-off (full objective in Appendix C). Because the Stage-I minimizer still factors as $q^*(x) = c\,\sigma_{\text{sim}}(x)$, this single trained network serves *all* coverage levels α and *all* deployment domains: every α-dependence lives in the simulator-computable constants (u, c), with $u = F_Z^{-1}(1-\alpha)$ and c the scalar proportionality constant induced by the full Stage-I objective (Appendix C). Conformalized quantile regression Romano et al. (2019) fits a *pair* of quantile regressors per level α.

Stage II — Scalar calibration. Normalize the real residuals, $V_i = R_i/q(X_i)$, and estimate one scalar $\hat{s}_b$ per stratum as the k_b-th order statistic of the within-stratum scores $\{V_i : X_i \in S_b\}$, $k_b = \lceil (n_b+1)(1-\alpha) \rceil$. This is split conformal on normalized scores Papadopoulos et al. (2002); Angelopoulos & Bates (2023), hence distribution-free regardless of q, h, or the simulator; what the simulator buys is that the *conditional* behavior is nontrivial. Stage II stores only K scalars atop one amortized network.

Stage III — Deployment. With no real labels ($n_{\text{cal}} = 0$), deploy the purely simulator-derived zero-shot threshold $s_P = u/c$; with real labels, deploy $\hat{s}_b$. A per-stratum certificate Cert_b (Section 4.4) provides a conservative audit of zero-shot coverage: it can retain zero-shot operation when the certified lower bound is sufficient, while inconclusive strata are flagged for scalar recalibration; a separate shape audit flags potential simulator mismatch. SUIT applies wherever a sampler stands in for the deployment environment and residuals approximately factorize into scale and shape; Section 5 instantiates two families — urn/occupancy models for unseen-species problems and neural posterior estimators in simulation-based inference (details in Appendix C).

4 Theory: Why Scale Calibration Works

This section formalizes the claim behind SUIT: under a multiplicative heteroscedastic error model, simulator training reduces the sim-to-real correction to the estimation of a scalar, and the residual coverage loss is governed by the *dispersion* of the log scale-ratio $\lambda(x) = \log\big(\sigma_{\text{real}}(x)/\sigma_{\text{sim}}(x)\big)$ rather than by its level, which Stage II absorbs exactly. The development follows a single chain — decompose the gap, remove its level, control the residual dispersion, and deploy. The zero-shot guarantee (Theorem 1) and the level–dispersion decomposition (Theorem 2 and Corollary 2, the theoretical center of the paper) are stated in full; all remaining results (Theorem 3, Propositions 1 and 2, Theorem 4 with Propositions 3 and 4, and Theorems 5–8) are summarized informally, with full statements and proofs in Appendix A.

4.1 Scale-gap decomposition

We work in the two-environment model of Section 2: $R = |Y - h(X)|$, $V = R/q(X)$, and $C(x; s) = [\, h(x) \pm s\, q(x) \,]$. Two structural assumptions underlie the whole development.

Assumption 1 (M: multiplicative heteroscedasticity)**.** In each environment $e \in \{\text{sim}, \text{real}\}$, $R = \sigma_e(X) \cdot Z_e$ with $\sigma_e : \mathcal{X} \to \mathbb{R}_+$, $Z_e \geq 0$, and $Z_e \perp X$ (P_e-a.s.).

Assumption 2 (S: shape preservation)**.** The shape law is shared, $\mathcal{L}_{\text{sim}}(Z) = \mathcal{L}_{\text{real}}(Z) =: \mathcal{L}_Z$; its cdf F_Z is strictly increasing with density f_Z bounded by $C_f := \sup_z f_Z(z) < \infty$. We write $u := F_Z^{-1}(1-\alpha)$.

The certificate and the end-to-end finite-sample bounds additionally require mild technical conditions — within-stratum sub-Gaussian dispersion of λ, bounded log-variables, non-degeneracy of f_Z near $1-\alpha$, and a pointwise Stage-I approximation bound; being theorem-specific, these are stated as Assumptions 3–6 in Appendix A, where the exact Stage-I case is also verified for the population minimizer of the Stage-I objective under Assumption 1.

A single algebraic identity drives every result of this section. Under Assumptions 1 and 2 with $q = c\,\sigma_{\text{sim}}$, the normalized score in the real environment factors as

$$V \;=\; \frac{\sigma_{\text{real}}(X)\,Z}{c\,\sigma_{\text{sim}}(X)} \;=\; \frac{\rho(X)\,Z}{c}, \qquad \text{hence} \qquad \mathbb{Q}\big(V \leq s \mid X = x\big) \;=\; F_Z\big(c\,s\,e^{-\lambda(x)}\big), \tag{4}$$

since $Z \perp X$. Thus normalization removes the high-dimensional heteroscedastic structure, leaving the entire sim-to-real gap encoded in the scalar field $\lambda(x)$ alone.

4.2 Scalar calibration removes the scale level

Zero-shot: the level hurts at first order. At $n_{\text{cal}} = 0$ the deployed threshold is the purely simulator-derived $s_P = u/c$, with coverage exactly $1-\alpha$ under $\mathbb{P}$. Write $\text{Cov}_{\text{real}}(s) := \mathbb{Q}(V \leq s)$ for the real-environment coverage of $C(\cdot\,; s)$, and $\text{Cov}_{\text{real}}(s \mid X = x)$ for its pointwise counterpart.

Theorem 1 (Zero-shot degradation bound)**.** *Under Assumptions 1 and 2 with $q = c\,\sigma_{\text{sim}}$, the zero-shot threshold $s_P = u/c$ satisfies*

$$\text{Cov}_{\text{real}}(s_P) \;\geq\; 1-\alpha \;-\; C_f\,u\,\mathbb{E}_X\Big[\,\big|e^{-\lambda(X)} - 1\big|\,\Big] \;\geq\; 1-\alpha \;-\; C_f\,u\,e^{\|\lambda\|_\infty}\,\mathbb{E}\big[\,|\lambda(X)|\,\big]. \tag{5}$$

The degradation is *first-order in the level* of λ: zero-shot calibration cannot absorb even a spatially constant scale error, so the simulator must get the noise magnitude itself right; under a faithful simulator, $\lambda \equiv 0$ and zero-shot coverage is exact.

Corollary 1 (Covariate-shift immunity)**.** *Under pure covariate shift with an unchanged conditional error model ($\sigma_{\text{sim}} = \sigma_{\text{real}}$, $\mathcal{L}_Z$ common, $Z \perp X$), zero-shot coverage remains exact, $\text{Cov}_{\text{real}}(s_P) = 1-\alpha$, even when the covariate supports are disjoint and the density ratio is unbounded or undefined, so that weighted-conformal guarantees Tibshirani et al. (2019) degenerate; the prerequisite is that q extrapolates structurally to the real support of X.*

Scalar calibration: only the dispersion remains. Now let Stage II receive n_{cal} i.i.d. real calibration points and set $\hat{s}$ to the k-th order statistic of $\{V_i\}$, $k = \lceil (n_{\text{cal}}+1)(1-\alpha) \rceil$. The central step is to decompose the gap into a *level* and a *dispersion*,

$$\lambda(x) \;=\; \bar{\lambda} + \epsilon(x), \tag{6}$$

where $\bar{\lambda} \in \mathbb{R}$ is any reference level of $\lambda(X)$ (e.g. its median under $\mathbb{Q}^X$) and $\epsilon(x) := \lambda(x) - \bar{\lambda}$. Let $G(t) := \mathbb{Q}(V \leq t) \;=\; \mathbb{E}_X\big[F_Z\big(c\,t\,e^{-\lambda(X)}\big)\big]$ denote the marginal score cdf. A standard finite-sample relaxation of split conformal Angelopoulos & Bates (2023) gives $G(\hat{s}) \geq 1-\alpha-\epsilon_n$ with probability $1-\beta$, $\epsilon_n = \sqrt{\ln(2/\beta)/(2n_{\text{cal}})} + O(1/n_{\text{cal}})$.

Theorem 2 (Level–dispersion decomposition)**.** *Under Assumptions 1 and 2 with $q = c\,\sigma_{\text{sim}}$, write $d(x) := e^{-\epsilon(x)}$ and $\tilde{t} := c\,\hat{s}\,e^{-\bar{\lambda}}$. Then with probability at least $1-\beta$ over the calibration set, simultaneously for every $x \in \mathcal{X}$,*

$$\text{Cov}_{\text{real}}(\hat{s} \mid X = x) \;\geq\; 1-\alpha \;-\; C_f\,\tilde{t}\,\Big(\,\big|e^{-\epsilon(x)} - 1\big| + \mathbb{E}_{X'}\big[\,\big|e^{-\epsilon(X')} - 1\big|\,\big]\,\Big) \;-\; \epsilon_n, \tag{7}$$

and a symmetric upper bound holds with the two dispersion terms added in absolute value. Also, under Assumption 5, $\tilde{t} = u \cdot \big(1 + O\big(\mathbb{E}|\epsilon| + \epsilon_n\big)\big)$ with explicit constants, so $C_f\,\tilde{t} = C_f\,u\,(1+o(1))$ whenever the dispersion is small.

Proof sketch. By equation 4, conditional coverage is $F_Z\big(\tilde{t}\,d(x)\big)$, whereas marginal calibration constrains its average over X; the Lipschitz bound on F_Z converts their difference into the two dispersion terms of equation 7, and inserting $G(\hat{s}) \geq 1-\alpha-\epsilon_n$ contributes the calibration error. Full proof in Appendix A.

The *level* $\bar{\lambda}$ appears only inside $\tilde{t}$, which the data-driven $\hat{s}$ pins to u; the coverage loss depends on λ only through its *dispersion* $\epsilon(x)$, plus the distribution-free calibration noise ϵ_n:

$$\text{coverage loss} \;\sim\; \underbrace{\text{dispersion of } \epsilon(x)}_{\text{irreducible}} \;+\; \underbrace{O\big(n_{\text{cal}}^{-1/2}\big)}_{\text{calibration noise}}. \tag{8}$$

Corollary 2 (Scalar absorption of arbitrary scale shift)**.** *If the sim-to-real gap is a pure global scale shift ($\sigma_{\mathrm{real}} = e^{\bar{\lambda}}\sigma_{\mathrm{sim}}$ with $\bar{\lambda} \in \mathbb{R}$ arbitrarily large, i.e. $\epsilon \equiv 0$), then $\mathrm{Cov}_{\mathrm{real}}(\hat{s} \mid X = x) \geq 1 - \alpha - \epsilon_n$ for every $x \in \mathcal{X}$, with $\epsilon_n = O\big(\sqrt{\log(1/\beta)/n_{\mathrm{cal}}}\big)$. That is, $O(\delta^{-2})$ real points estimating a* single scalar *restore pointwise near-conditional coverage under scale misspecification of arbitrary magnitude (n_{cal} in the low hundreds yields $\epsilon_n \approx 0.07$).*

Takeaway. Zero-shot coverage depends on the full scale-ratio *level*, whereas scalar calibration removes any global scale shift and leaves only within-stratum heterogeneity plus finite-sample calibration error: Stage II purchases exactly the one dimension of the gap, scale level, that normalization by q cannot supply, at a distribution-free rate. Shape misspecification degrades zero-shot validity and efficiency but not the distribution-free validity of real-data calibration (Theorem 3, Appendix A), a regime the audit of Proposition 4 detects. Simpler routes fail quantitatively: no global raw-score threshold repairs shape heterogeneity at any sample size (Proposition 1, Appendix A), and weighted conformal prediction Tibshirani et al. (2019) requires an estimable density ratio and restores only marginal coverage with a global scalar (Proposition 2, Appendix A).

4.3 Stratification controls the residual heterogeneity

A single scalar removes only the average scale shift; stratification reduces the remaining within-stratum dispersion, at the price of thinner per-stratum calibration sets. Within each stratum, split conformal remains marginally valid by exchangeability alone, exactly on average at any K (Theorem 5, Appendix A). The realized per-stratum guarantee, however, incurs simultaneous calibration noise

$$\epsilon_{n_b} \;=\; O\Big(\sqrt{\log(K/\beta)/n_b}\Big) \tag{9}$$

across all K strata, and realized slack δ everywhere is certified by $n_{\mathrm{cal}} = O\big(K \log K \,/\, \delta^2\big)$ under balanced allocation (Theorem 6, Appendix A); the $\log K$ is the price of simultaneous attainment, the bottleneck is the thinnest stratum, and when $n_b + 1 < 1/\alpha$ the realized guarantee is vacuous: it speaks where calibration cannot yet. Balancing the dispersion term, which scales as $O(1/K)$, against this calibration variance gives

$$K^* \;=\; \Theta\Big(\big(n_{\mathrm{cal}}/\log n_{\mathrm{cal}}\big)^{1/3}\Big)\,, \tag{10}$$

the optimum of the constructed upper bound (*bound-optimal stratification*, Theorem 7, Appendix A): without stratification ($K = 1$) the worst-case pointwise loss entering the bound is $\Theta(1)$, while bound-optimal stratification compresses it to $\widetilde{O}\big(n_{\mathrm{cal}}^{-1/3}\big)$.

4.4 Deployment audit and the end-to-end bound

Bounds involving total variation or density ratios Foygel Barber et al. (2023); Tibshirani et al. (2019) are generally unobservable at deployment; we instead identify directly the object that determines coverage. From equation 4,

$$\log V \;=\; \lambda(X) + W, \qquad W := \log(Z/c)\,: \tag{11}$$

the distribution of W is simulator-computable (its moments $m_1 := \mathbb{E}[W]$, $m_2 := \mathrm{Var}(W)$), and $\lambda(X) \perp W$, so a few real audit points identify the first two per-stratum moments of $\lambda(X)$ by moment subtraction. Theorem 4 (Appendix A) turns this deconvolution into a per-stratum certificate: with probability at least $1 - 2K\beta'$,

$$\boxed{\mathrm{Cov}_b(s_P) \;\geq\; \mathrm{Cert}_b} \qquad \text{simultaneously over } b = 1, \ldots, K, \tag{12}$$

where $\mathrm{Cov}_b(s_P) := \mathbb{Q}\big(V \leq s_P \,\big|\, X \in S_b\big)$ is the stratum-b zero-shot coverage and Cert_b is computable from simulator quantities and as few as $n_b \geq 2$ real audit points, up to two a-priori constants: an analytic bound C_f on the simulator density peak and a range bound M_λ on the log scale-ratio (Assumptions 3–4, Appendix A). Cert_b is asymptotically zero-tight and consistent at rate $O(n_b^{-1/2})$, and a Kolmogorov–Smirnov shape audit accompanies it (Proposition 4, Appendix A). The certificate is a conservative sufficient test: a positive value certifies that much stratum coverage, while a vacuous value is inconclusive and triggers recalibration; the full construction, concentration radii, and audit are deferred to Appendix A.

Combining Stage-I approximation, within-stratum scale dispersion, and finite-sample calibration gives the end-to-end bound (Theorem 8, Appendix A): with probability at least $1 - \beta$ over the calibration set, simultaneously for every x (with $b = b(x)$),

$$\mathrm{Cov}_{\mathrm{real}}(\hat{s}_b \mid x) \;\geq\; 1 - \alpha \;-\; \underbrace{\Delta_b(x)}_{\text{Stage-I + within-stratum dispersion}} \;-\; \underbrace{\epsilon_{n_b}}_{\text{calibration}}\,, \tag{13}$$

where $\Delta_b(x)$ is the effective-dispersion term of Theorem 8: the scale-ratio *level* is absorbed by $\hat{s}_b$ (Corollary 2), and the Stage-I pointwise error is folded into the within-stratum dispersion alongside the scale-ratio heterogeneity. With

bound-optimal stratification equation 10, the calibration component is $\widetilde{O}\big(n_{\mathrm{cal}}^{-1/3}\big)$. Theorem 8 is the sufficient-data end of a spectrum indexed by the per-stratum real sample size: at $n_b = 0$ the zero-shot bound of Theorem 1 applies, for $2 \le n_b \lesssim \delta^{-2} \ln K$ the certificate Cert_b speaks where realized guarantees are still vacuous, and under the audit-then-calibrate protocol of Section 3.2 these compose by union bound into a single end-to-end inequality with an explicit, allocatable confidence budget.

5 EXPERIMENTS

The experiments are organized around the three questions. (Q1) Does SUIT deliver accurate, adaptive intervals with zero or few real labels (Section 5.2)? (Q2) *Why* does the two-stage decomposition work—is the sim-to-real gap really one-dimensional (Section 5.3)? (Q3) When the simulator cannot be trusted, can we detect that zero-shot deployment is unsafe (Section 5.4)?

5.1 EXPERIMENTAL SETUP

Datasets. *(i) NDV*: Seven real-world datasets that have been widely adopted in prior studies on distinct value estimation Wu et al. (2021); Li et al. (2024); Zhang et al. (2026): Kasandr Sidana et al. (2017), Airline of Transportation Statistics BTS (2020), DMV of Motor Vehicles NYS DMV (2020), Campaign FEC (2020), SSB O'Neil et al. (2009), NCVoter of Elections NCSBE (2020), and Census UCI Repository (2001). Bernoulli-sampled at rate 0.5%; h is a learned NDV estimator trained on urn-model simulations (the full scenario and pipeline are described in Appendix D). *(ii) Species richness*: the BCI tree census (21,457 individuals, 225 species) and GBIF Taiwan occurrence counts (20.5M records, top-1000 species; singletons truncated by construction), with Chao1 Chao (1984) as h; subsample sizes $n \in \{100, 250, 500, 1000, 2000\}$. *(iii) SBI*: Two-Moons Tejero-Cantero et al. (2020) with an NPE posterior median as h; the "real" environment injects extra noise of scale $\rho \in \{0, 1, 1.5, 2\}$, Gaussian or Student-$t$. *(iv) Synthetic regression*: $X \sim \mathcal{N}(0, I_8)$, gradient-boosted h.

Baselines. SCP (split CP Papadopoulos et al. (2002); Lei et al. (2018)) and SSCP (split CP applied separately per tail-index stratum, without the learned scale; Appendix D)—valid at any budget, but the threshold is global, respectively piecewise-constant per stratum. *Target-trained adaptive*: LCP Guan (2023); CQRD, CQR Romano et al. (2019) refit per α on the n_{cal} calibration points themselves; CQRF, a full-data CQR reference—adaptive, but the shape is paid for with target labels. *Simulator- and shift-based*: WCP, weighted CP with an estimated density ratio Tibshirani et al. (2019), which corrects shift but keeps a scalar threshold; CQRS, CQR trained in the simulator and calibrated on real data, the strongest classical competitor; CQSS, CQR trained *and* calibrated in the simulator (the simulator-internal strategy of CP4SBI Cabezas et al. (2025)), which needs no real labels but leaves the sim-to-real gap uncontrolled. SUIT itself runs zero-shot (SUIT-Z), with one calibrated scalar (SUIT-G), or with per-stratum scalars (SUIT). All experiments target $1-\alpha = 0.95$; the few-shot budget is $n_{\mathrm{cal}} = 10\%$ of columns (NDV) or 100 points (other domains). Results are 5-run means; error bars are omitted as $SD \le 0.01$.

5.2 MAIN RESULTS: ACCURATE INTERVALS WITH ZERO OR FEW REAL LABELS

Tables 2 and 3 report results per dataset at the two ends of the calibration spectrum. Three findings summarize the comparison.

(1) Zero-shot: when the simulator is faithful, no real data are needed—and no baseline offers this operating point. At $n_{\mathrm{cal}} = 0$ every calibration-dependent method is undefined. SUIT-Z covers at least 0.90 on all seven NDV databases at relative width 1.1–2.8, 0.95 on Two-Moons at width 0.83, and 0.95 on the synthetic domain; on species it is valid everywhere (coverage 1.00) at widths 3.7–12.5. The only simulator-only competitor, CQSS, either covers vacuously (relative widths of $6{\times}10^6$–$5.2{\times}10^7$ on NDV) or collapses to zero coverage on GBIF at $n = 2000$. Under injected noise shift ($\rho \ge 1$) both zero-shot methods collapse below 0.12—the regime Section 5.4 investigates.

(2) Few-shot on NDV: a 10% calibration-column budget replaces the missing scale. At $n_{\mathrm{cal}} = 10\%$ of the columns, SUIT covers 0.99–1.00 at relative width 1.1–5.4. Every competitor either under-covers—CQRD at 0.00–0.24, since so few labeled columns cannot fit two quantile heads; WCP at 0.40–0.63 on six of seven databases; SSCP at 0.50–1.00, its per-bin quantiles degenerate at this budget—or buys coverage with widths four to five orders of magnitude larger (SCP, LCP, CQRF, CQRS at $4{\times}10^4$–$5{\times}10^5$). Stratification is load-bearing: with one global scalar, SUIT-G is up to $3.3\times$ wider (Census 3.74 vs. 1.15; NCVoter 5.05 vs. 1.69).

(3) Where adaptivity is cheap, the gap closes—we report this as-is. On species, SSCP, LCP and the SUIT variants are all width-comparable (widths 0.36–1.9); the coverage floors differ: SSCP 0.97, SUIT 0.98, LCP 0.83. On the synthetic domain the SUIT variants are the narrowest near-valid methods (5.9–7.6 at coverage 0.93–0.94), ahead of SSCP (9.7) and SCP/WCP (16.9). Under injected SBI shift all calibrated methods stay tight (widths 4.5–17.6): stratified SUIT holds coverage 0.98–0.99 at width 4.5–4.6, comparable to SSCP (0.98–0.99, 5.2–5.5) and CQRS (0.95, 4.7–4.9).

Table 2: Few-shot per-dataset results (target coverage $1 - \alpha = 0.95$).

Domain	Dataset		NDV Airlines	NDV Campaign	NDV Census	NDV DMV	NDV NCVoter	NDV SSB	NDV Kasandr	Species BCI n=100	Species BCI n=2000	Species GBIF-TW n=100	Species GBIF-TW n=2000	SBI Two-Moons ρ=0	SBI Two-Moons ρ=2, G	SBI Two-Moons ρ=2, t	Synth Synth nominal
Conformal	SCP	cov.	1.00	1.00	1.00	1.00	0.96	0.71	0.86	0.91	0.98	0.99	0.98	0.98	0.95	0.95	0.93
		rel. width	4.3×10^4	1.3×10^5	3.8×10^5	1.4×10^5	1.4×10^5	1.4×10^5	3.1×10^5	1.36	0.49	1.76	0.92	0.91	4.69	4.91	16.9
	SSCP	cov.	0.50	1.00	1.00	0.80	0.75	0.59	0.57	0.97	0.98	0.99	0.98	0.98	0.98	0.99	0.92
		rel. width	2.96	1.0×10^4	26.2	836.8	7.4×10^3	10.2	22.3	1.48	0.42	1.71	0.86	1.05	5.25	5.48	9.72
Target-trained	LCP	cov.	1.00	1.00	1.00	1.00	0.96	0.71	0.86	0.83	0.92	0.94	0.95	0.95	0.94	0.93	0.85
		rel. width	4.3×10^4	1.3×10^5	3.8×10^5	1.4×10^5	1.4×10^5	1.4×10^5	3.1×10^5	1.23	**0.36**	1.70	**0.81**	0.89	4.81	4.79	9.04
	CQRD	cov.	0.20	0.05	0.17	0.05	0.10	0.24	0.00	0.97	0.99	1.00	0.96	0.99	1.00	1.00	0.95
		rel. width	0.17	1.01	1.64	0.70	0.78	0.60	2.15	1.95	1.93	1.99	1.98	1.17	9.30	17.6	44.1
	CQRF	cov.	1.00	0.80	0.98	0.95	0.93	0.71	0.86	0.89	0.98	0.96	0.96	0.99	1.00	1.00	0.91
		rel. width	5.4×10^4	1.6×10^5	4.8×10^5	1.8×10^5	1.7×10^5	1.8×10^5	3.9×10^5	1.96	1.89	1.99	1.98	1.15	9.82	11.5	34.9
Sim. / shift	WCP	cov.	0.40	0.45	1.00	0.60	0.63	0.47	0.43	0.91	0.98	0.90	0.98	0.95	1.00	0.99	0.93
		rel. width	2.96	8.70	26.2	9.68	9.38	9.93	21.2	1.35	0.49	1.72	0.93	0.81	6.13	6.12	16.9
	CQRS	cov.	1.00	0.80	0.98	0.95	0.93	0.71	0.86	0.97	0.91	0.99	0.95	0.98	0.95	0.95	0.94
		rel. width	5.4×10^4	1.6×10^5	4.8×10^5	1.8×10^5	1.7×10^5	1.8×10^5	3.9×10^5	20.8	23.9	3.37	5.25	1.02	4.70	4.92	44.5
SUIT	SUIT-G	cov.	1.00	0.95	1.00	1.00	0.99	1.00	0.96	0.95	0.98	0.98	0.98	0.98	0.94	0.95	0.94
		rel. width	3.06	6.15	3.74	5.55	5.05	5.92	5.00	1.52	0.62	1.93	0.97	0.95	4.80	5.05	**5.85**
	SUIT	cov.	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	1.00	0.98	0.99	0.98	0.99	0.93
		rel. width	**1.06**	**4.25**	**1.15**	**5.43**	**1.69**	**3.82**	**2.73**	**1.30**	0.46	**1.67**	0.84	**0.80**	**4.53**	**4.61**	7.58

Legend. Calibration budget: $n_{\text{cal}} = 10\%$ of columns (NDV) or 100 points (other domains). Per method, empirical coverage (cov.) and mean relative width (rel. width; interval width relative to the true value—SBI reports absolute width, its target residual being 0) appear in separate rows. Underlined coverage falls below 0.90; **bold** marks the smallest width among methods reaching coverage ≥ 0.90 in that column. On NDV, LCP's neighbourhood spans all calibration columns, so it coincides numerically with SCP.

Table 3: Zero-shot results ($n_{\text{cal}} = 0$); Calibration-dependent methods fail, leaving only SUIT-Z and CQSS.

Domain	Dataset		NDV Airlines	NDV Campaign	NDV Census	NDV DMV	NDV NCVoter	NDV SSB	NDV Kasandr	Species BCI n=100	Species BCI n=2000	Species GBIF-TW n=100	Species GBIF-TW n=2000	SBI Two-Moons ρ=0	SBI Two-Moons ρ=2, G	SBI Two-Moons ρ=2, t	Synth Synth nominal
CQSS		cov.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.00	0.98	0.04	0.04	0.94
		rel. width	6.0×10^6	1.7×10^7	5.2×10^7	1.9×10^7	2.1×10^7	2.0×10^7	4.3×10^7	28.1	25.1	5.92	4.03	1.01	0.96	0.95	44.5
SUIT-Z		cov.	1.00	0.91	1.00	0.95	0.90	1.00	0.93	1.00	1.00	1.00	1.00	0.95	0.02	0.01	0.95
		rel. width	**1.13**	**1.44**	**1.32**	**2.41**	**1.72**	**2.76**	**1.93**	**12.5**	**11.6**	**3.67**	**5.05**	**0.83**	0.67	0.66	**7.14**

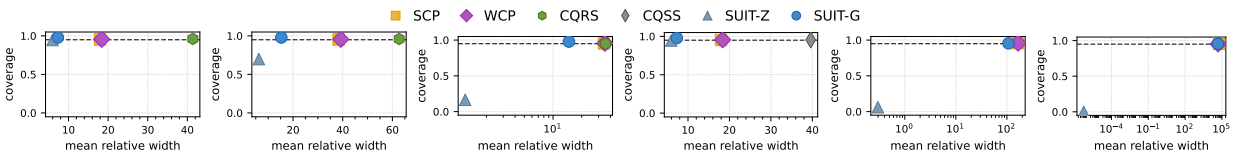
5.3 Why Does Scalar Calibration Work?

The main result shows where SUIT stands; this section shows *why*, using the synthetic domain where each mismatch axis can be turned independently, and per-stratum coverage on the urn domains where heteroscedasticity lives.

Zero-shot pays for the entire scale mismatch; one scalar removes it. Figure 3(a–c) multiplies the real noise scale by $\rho \in \{1, 2, 10\}$ at fixed shape and covariates, one panel per ρ. Zero-shot coverage absorbs the whole mismatch—0.95, 0.70, 0.16 from (a) to (c)—while its width stays frozen at the nominal level: in the coverage–width plane SUIT-Z slides down and to the left, narrower but increasingly invalid. One hundred real points estimating a single scalar hold coverage at 0.98 at every ρ—the calibrated methods stay pinned at the target line while their widths track the inflation; at $\rho = 10$, SUIT-G is $2.5\times$ narrower than the strongest baseline (relative width 13.8 vs. 35.5 for CQRS; SCP and WCP ≈ 35). The scalar removes the level mismatch; the learned shape is what keeps the interval narrow.

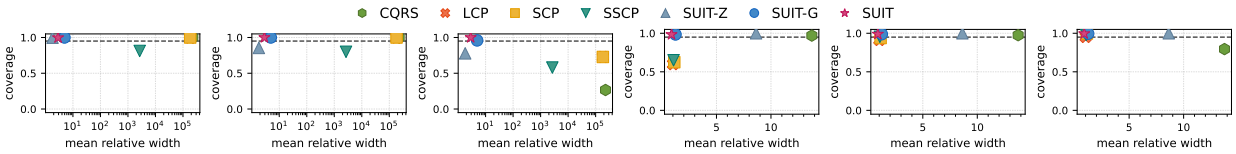
Global calibration cannot fix heteroscedasticity; the learned shape can. Figure 4 reports conditional coverage per tail stratum on the urn domains. On NDV every baseline collapses in the heavy stratum—SCP and LCP 0.73, SSCP 0.58, CQRS 0.27—and zero-shot SUIT-Z degrades to 0.78; one global scalar (SUIT-G) repairs most of the deficit (0.88), and only the per-stratum variant SUIT holds all three strata at 1.00 (Figure 4(c)). On species the light stratum is data-starved for every urn-calibrated method (0.60–0.76 at 10–30 test points per stratum) while SUIT holds the mid and heavy strata at 0.95/0.98; CQRS shows the opposite profile (0.97 light, 0.79 heavy)—it learned where the light queries are, not the heavy ones.

Covariate shift is a separate axis—and it breaks zero-shot extrapolation. Figure 3(d–f) shifts only the covariate marginal until the supports are disjoint, one panel per μ. SUIT-Z collapses to 0.06–0.00 for $\mu \geq 3$: q misextrapolates off the simulated support, the same failure as simulator-only CQSS (0.00). Calibration survives but pays for the extrapolation error in width: SUIT-G, SCP and WCP all hold 0.95–0.96 at relative widths of order 5×10^4 at $\mu = 6$, SUIT-G narrowest (4.8×10^4); WCP coincides numerically with SCP, its density-ratio estimate saturating under disjoint supports. This panel is the empirical motivation for the deployment monitor of Section 5.4.



(a) ρ=1 (b) ρ=2 (c) ρ=10 (d) μ=0 (e) μ=3 (f) μ=6

Figure 3: Synthetic separation experiments: coverage versus relative width under pure scale shift $\sigma_{\text{real}} = \rho\,\sigma_{\text{sim}}$ (a–c) and pure covariate shift μ with disjoint supports (d–f).



(a) NDV, light (b) NDV, mid (c) NDV, heavy (d) Species, light (e) Species, mid (f) Species, heavy

Figure 4: Per-stratum conditional coverage versus mean relative width.

5.4 Deployment: Detecting When Zero-Shot Is Unsafe

Theorem 4 makes Cert_b a candidate deployment instrument with three testable properties: validity (it lower-bounds true stratum coverage), zero-tightness (no alarm when the simulator is right), and monotone tracking (it degrades as the gap grows). The deployed certificate implements equation 19 with two validity-preserving tightenings: a *local-Lipschitz* correction replacing the global sup-norm C_f by the maximum of f_Z over the effective interval $[u\,e^{-\bar{\lambda}-3\hat{\sigma}},\ u\,e^{-\bar{\lambda}}]$ (simulator-computable), and a *convexity-verified Jensen step* (Proposition 3): when $g(t) = F_Z(u\,e^{-t})$ is numerically convex on the moment-estimated support of λ (criterion $(z\,f_Z(z))' \geq 0$), the entire σ-correction vanishes and $\text{Cert}_b = F_Z(u\,e^{-(\hat{\lambda}+\delta_\lambda)})$; otherwise the bound falls back to equation 19. We stress-test along the mismatch path (ρ, σ_λ) from $(1, 0)$ to $(10, 0.5)$, audit sizes $n_b \in \{50, 500, 2000, 5000\}$, 200 repetitions per cell (Figure 5). Panel (a) plots true zero-shot coverage (black) and the certificate at n_b=5000 (red, p10–p90 band) against $\bar{\lambda} = \log\rho$; panel (b) plots the certificate versus audit size n_b at the three leftmost path points, with true coverage dotted.

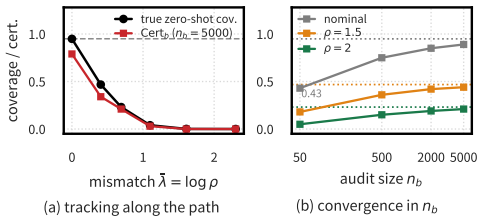


Figure 5: Certificate stress test.

All three properties hold. *Validity*: the violation rate is 0 in all cells, including those where the Jensen step fires. *Tracking*: the certificate is non-trivial exactly in the small-to-moderate mismatch regime and degrades monotonically with the gap while converging upward in n_b toward the true coverage—$0.18/0.36/0.42/0.44$ against a true 0.468 at $(\rho, \sigma_\lambda) = (1.5, 0.05)$, and $0.05/0.15/0.19/0.21$ against 0.231 at $(2, 0.1)$. At large mismatch the certificate reads 0—by design: the instrument only needs to speak when zero-shot is nearly safe, and a vacuous reading there correctly triggers recalibration. Operationally the certificate is one-sided evidence: a positive value certifies that much stratum coverage; a zero reading carries no information and routes to the realized-coverage audit of Proposition 4.

Zero-tightness holds, by noise-family design. The structural obstruction of Remark 9(b)—any density that decreases past its $(1-\alpha)$-quantile fails the convexity check on an interval straddling it—is avoided by simulating with a bounded noise family whose density increases throughout its support (Beta-type): then $z\,f_Z(z)$ is nondecreasing on the whole support, the Jensen step fires at the nominal point for every audit size, and the σ-correction vanishes exactly where the certificate must be tight. At zero mismatch the certificate reads $0.43/0.75/0.85/0.89$ at $n_b = 50/500/2000/5000$ against a true coverage of 0.949, concentrating on $F_Z(u) = 1 - \alpha$ at the rate of δ_λ alone (Remark 9(a)).

6 Conclusion

We presented SUIT, a simulator-trained conformal framework that separates uncertainty into a transferable shape and a real-data-calibrated scalar. This decomposition removes global sim-to-real scale shifts while stratification controls residual heterogeneity, enabling adaptive intervals with zero or few real labels. Experiments across four data families show near-nominal zero-shot coverage under faithful simulation and recovery under substantial scale mismatch with limited calibration data.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and grammar correction of the manuscript. No generative AI tools were used for any tasks beyond the language editing described above. All AI-suggested edits were limited to wording and grammar; each revised sentence was manually checked by the authors to ensure that the original technical meaning, claims, and numerical results were fully preserved. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All theoretical claims are proved in the Appendix. Anonymous source code is available at `https://anonymous.4open.science/r/SUIT-E8B6`.

REFERENCES

Shachar Angelman, Rotem Nizhar, and Jacob Goldberger. Calibrating without labels: Source-free conformal prediction using pseudo-labels. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications (COPA)*, volume 266 of *Proceedings of Machine Learning Research*, pp. 63–81. PMLR, 2025.

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

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://arxiv.org/abs/2208.02814`.

Liviu Aolaritei, Julie Zhu, Oliver Wang, Michael Jordan, and Youssef Marzouk. Conformal prediction under lévy–prokhorov distribution shifts: Robustness to local and global perturbations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL `https://arxiv.org/abs/2502.14105`.

Anna Ben-Hamou, Stéphane Boucheron, and Mesrob I. Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1):249–287, 2017.

Luben M. C. Cabezas, Vagner S. Santos, Thiago R. Ramos, Pedro L. C. Rodrigues, and Rafael Izbicki. CP4SBI: Local conformal calibration of credible sets in simulation-based inference. arXiv:2508.17077, 2025. URL `https://arxiv.org/abs/2508.17077`.

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

Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11(4):265–270, 1984.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and André F. T. Martins. Non-exchangeable conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024. URL `https://arxiv.org/abs/2310.01262`.

Federal Election Commission FEC. Campaign finance data. `https://www.fec.gov/data/`, 2020.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Few-shot conformal prediction with auxiliary tasks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

Jonas Flechsig and Maximilian Pilz. Conformal prediction without knowledge of labeled calibration data. arXiv:2509.10321, 2025. URL `https://arxiv.org/abs/2509.10321`.

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.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Irving J. Good and George H. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

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

Samuel Karlin. Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, 17 (4):373–401, 1967.

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

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.

Benjamin LeRoy and David Zhao. MD-split+: Practical local conformal inference in high dimensions, 2021. URL https://arxiv.org/abs/2107.03280. ICML 2021 Workshop on Distribution-Free Uncertainty Quantification; arXiv:2107.03280.

Jiajun Li, Runlin Lei, Sibo Wang, Zhewei Wei, and Bolin Ding. Learning-based property estimation with polynomials. *Proc. ACM Manag. Data*, 2(3):148, 2024. doi: 10.1145/3654994. URL https://doi.org/10.1145/3654994.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pp. 343–351. PMLR, 2021.

Hui Ma and Kai Yang. MetaSTNet: Multimodal meta-learning for cellular traffic conformal prediction. arXiv:2505.21553, 2025. URL https://arxiv.org/abs/2505.21553.

North Carolina State Board of Elections NCSBE. Voter registration statistics. https://www.ncsbe.gov/results-data/voter-registration-data, 2020.

New York State Department of Motor Vehicles NYS DMV. Vehicle, snowmobile, and boat registrations. https://catalog.data.gov/dataset/vehicle-snowmobile-and-boat-registrations, 2020.

Bureau of Transportation Statistics BTS. Airlines departure delay. https://www.openml.org/d/42728, 2020.

Patrick E. O'Neil, Elizabeth J. O'Neil, Xuedong Chen, and Stephen Revilak. The star schema benchmark and augmented fact table indexing. In Raghunath Othayoth Nambiar and Meikel Poess (eds.), *Performance Evaluation and Benchmarking, First TPC Technology Conference, TPCTC 2009, Lyon, France, August 24-28, 2009, Revised Selected Papers*, volume 5895 of *Lecture Notes in Computer Science*, pp. 237–252. Springer, 2009. doi: 10.1007/978-3-642-10424-4_17. URL https://doi.org/10.1007/978-3-642-10424-4_17.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning (ECML)*, pp. 345–356. Springer, 2002.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. PAC prediction sets for meta-learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 37920–37931, 2022.

Yash Patel, Declan McNamara, Jackson Loper, Jeffrey Regier, and Ambuj Tewari. Variational inference with coverage guarantees in simulation-based inference, 2023. URL https://arxiv.org/abs/2305.14275. CANVI; arXiv:2305.14275.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence (UAI)*, volume 161 of *Proceedings of Machine Learning Research*, pp. 844–853. PMLR, 2021.

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

Kaizer Rahaman, Jyotirmoy V. Deshmukh, Ashish R. Hota, and Lars Lindemann. When environments shift: Safe planning with generative priors and robust conformal prediction. arXiv:2602.12616, 2026. URL https://arxiv.org/abs/2602.12616.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Matteo Sesia, Stefano Favaro, and Edgar Dobriban. Conformal frequency estimation using discrete sketched data with coverage for distinct queries. *Journal of Machine Learning Research*, 24(348):1–80, 2023.

Farbod Siahkali, Ashwin Verma, and Vijay Gupta. Coverage guarantees for pseudo-calibrated conformal prediction under distribution shift. arXiv:2602.14913, 2026. URL https://arxiv.org/abs/2602.14913.

Sumit Sidana, Charlotte Laclau, Massih-Reza Amini, Gilles Vandelle, and André Bois-Crettez. KASANDR: A large-scale dataset with implicit feedback for recommendation. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (eds.), *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pp. 1245–1248. ACM, 2017. doi: 10.1145/3077136.3080713. URL https://doi.org/10.1145/3077136.3080713.

Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J. Gonçalves, David S. Greenberg, and Jakob H. Macke. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020.

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, 2019.

UCI Machine Learning Repository UCI Repository. US Census Data (1990). https://doi.org/10.24432/C5VP42, 2001.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Renzhi Wu, Bolin Ding, Xu Chu, Zhewei Wei, Xiening Dai, Tao Guan, and Jingren Zhou. Learning to be a statistician: Learned estimator for number of distinct values. *Proc. VLDB Endow.*, 15(2):272–284, 2021. doi: 10.14778/3489496.3489508. URL http://www.vldb.org/pvldb/vol15/p272-wu.pdf.

Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 2019.

Rui Xu, Chao Chen, Yue Sun, Parvathinathan Venkitasubramaniam, and Sihong Xie. Wasserstein-regularized conformal prediction under general distribution shift. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2501.13430.

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

Yujie Zhang, Xiaochun Yang, Bin Wang, and Yuan Sui. From single to multiple attributes: Experimental insights on sampling-based distinct combination estimation in group-by queries. In *42nd IEEE International Conference on Data Engineering, ICDE 2026, Montreal, QC, Canada, May 4-8, 2026*, pp. 1589–1602. IEEE, 2026. doi: 10.1109/ICDE65706.2026.00122. URL https://doi.org/10.1109/ICDE65706.2026.00122.

A Auxiliary Concentration Tools

This appendix collects the complete proofs of every result in Section 4: Theorems 1, 2, 3, 4, 5, 6, 7, and 8, Corollaries 1 and 2, and Propositions 1, 2, 4, and 3 (the convexity-verified Jensen variant of Theorem 4). The main text states Theorems 1 and 2 (with Corollaries 1 and 2) in full and summarizes all remaining results informally; their full formal statements — including the computable certificate equation 19, which the main text presents only through the guarantee $\mathrm{Cov}_b(s_P) \geq \mathrm{Cert}_b$ — are given here, immediately before their proofs. All notation, assumption labels, and theorem numbers refer to Section 4; conditioning on a stratum index "$\mid b$" always means conditioning on the event $X \in S_b$. Throughout, $\mathrm{TV}(\mu,\nu) := \sup_A |\mu(A) - \nu(A)|$ denotes total variation with the event-supremum convention, and $\mathrm{Beta}(a,b)$, $\mathrm{Bin}(n,q)$ denote the Beta and Binomial laws. The present section states the concentration tools used repeatedly below, each with explicit constants and applicability conditions.

Lemma 1 (Pushforward contraction of total variation)**.** *Let μ, ν be probability measures on a measurable space and ϕ a measurable map into a second measurable space. Then*

$$\mathrm{TV}\big(\mu \circ \phi^{-1},\, \nu \circ \phi^{-1}\big) \;\leq\; \mathrm{TV}(\mu,\nu).$$

Consequently, for any threshold s determined independently of the real data (e.g. simulator-determined),

$$\big|\mathrm{Cov}_{\mathrm{real}}(s) - \mathrm{Cov}_{\mathrm{sim}}(s)\big| \;\le\; \mathrm{TV}\big(P^{V}_{\mathrm{sim}},\, P^{V}_{\mathrm{real}}\big) \;\le\; \mathrm{TV}\big(P^{X,Y}_{\mathrm{sim}},\, P^{X,Y}_{\mathrm{real}}\big),$$

where P^{V}_{e} is the one-dimensional law of $V = R/q(X)$ under environment e.

Proof. With the event-supremum convention, the dual representation is $\mathrm{TV}(\mu,\nu) = \frac{1}{2}\sup_{\|g\|_\infty \le 1} \big|\int g\, d\mu - \int g\, d\nu\big|$, the supremum over bounded measurable g. For any such g, $g\circ\phi$ is again bounded measurable with $\|g\circ\phi\|_\infty \le 1$, and $\int g\, d(\mu\circ\phi^{-1}) = \int (g\circ\phi)\, d\mu$; hence the supremum defining the pushforward TV ranges over a subset of the functions defining $\mathrm{TV}(\mu,\nu)$, giving the first claim. For the second claim, the coverage event $\{Y \in C(X;s)\} = \{V \le s\}$ is an event of V alone, so $\big|\mathrm{Cov}_{\mathrm{real}}(s) - \mathrm{Cov}_{\mathrm{sim}}(s)\big| \le \mathrm{TV}(P^{V}_{\mathrm{sim}}, P^{V}_{\mathrm{real}})$ by the definition of total variation; applying the first claim with $\phi(X,Y) = |Y - h(X)|/q(X)$ gives the rightmost inequality. □

Lemma 2 (Hoeffding)**.** *Let $Y_1,\dots,Y_n$ be independent with $a \le Y_i \le b$ almost surely, $R := b - a$, and $\bar{Y} := n^{-1}\sum_i Y_i$. Then for every $t > 0$,*

$$\mathbb{P}\big(\big|\bar{Y} - \mathbb{E}\bar{Y}\big| \ge t\big) \;\le\; 2\exp\Big(-\frac{2nt^2}{R^2}\Big).$$

Equivalently, $\big|\bar{Y} - \mathbb{E}\bar{Y}\big| \le R\sqrt{\ln(2/\beta)/(2n)}$ with probability at least $1-\beta$. Applicability requires only independence and almost-sure boundedness; no further distributional structure is used.

Proof. Standard exponential-mgf argument: Hoeffding's lemma gives $\mathbb{E}e^{s(Y_i - \mathbb{E}Y_i)} \le e^{s^2R^2/8}$, and Markov's inequality applied to $e^{\pm s\sum_i (Y_i - \mathbb{E}Y_i)}$ with $s = 4t/R^2$ yields each one-sided bound $\le e^{-2nt^2/R^2}$. See any standard reference (e.g. the statement collected in Angelopoulos & Bates (2023)). □

Lemma 3 (McDiarmid; application to the U-statistic sample variance)**.** (a) Bounded differences. *Let $Y_1,\dots,Y_n$ be independent and let f satisfy the bounded-difference condition: replacing Y_i by an arbitrary y_i' changes f by at most c_i. Then for every $t > 0$,*

$$\mathbb{P}\big(\big|f - \mathbb{E}f\big| \ge t\big) \;\le\; 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big).$$

(b) Sample variance. *Suppose additionally $a \le Y_i \le b$ almost surely, $R := b - a$, and let*

$$s^2 \;:=\; \frac{1}{2n(n-1)}\sum_{i,j=1}^{n}(Y_i - Y_j)^2$$

be the unbiased sample variance (the sum runs over ordered pairs; s^2 coincides with the usual $(n-1)^{-1}\sum_i (Y_i - \bar{Y})^2$ and satisfies $\mathbb{E}s^2 = \mathrm{Var}(Y_1)$). Then the bounded-difference constants may be taken as $c_i = R^2/n$, and consequently

$$\mathbb{P}\big(\big|s^2 - \mathrm{Var}(Y_1)\big| \ge t\big) \;\le\; 2\exp\Big(-\frac{2nt^2}{R^4}\Big), \qquad \textit{i.e.} \qquad \big|s^2 - \mathrm{Var}(Y_1)\big| \le R^2\sqrt{\tfrac{\ln(2/\beta)}{2n}}$$

with probability at least $1-\beta$. Applicability requires independence and almost-sure boundedness; unbiasedness (hence the absence of a bias term) is the reason for the U-statistic form: a $1/n$ plug-in variance would add an $O(R^2/n)$ bias.

Proof. (a) is McDiarmid's inequality, standard. (b) The identity $s^2 = (n-1)^{-1}\sum_i (Y_i - \bar{Y})^2$ and $\mathbb{E}s^2 = \mathrm{Var}(Y_1)$ are classical. For the bounded differences: changing Y_i to Y_i' affects exactly the $2(n-1)$ ordered pairs containing index i; each affected summand $(Y_i - Y_j)^2 \in [0, R^2]$ changes by at most R^2, so the total change in s^2 is at most $2(n-1)R^2/\big(2n(n-1)\big) = R^2/n$. Part (a) with $\sum_i c_i^2 = n\cdot R^4/n^2 = R^4/n$ gives the claim. □

Lemma 4 (Beta–binomial identity and order-statistic lower tail)**.** *Let $Y_1,\dots,Y_n$ be i.i.d. with* continuous *cdf G, let $Y_{(k)}$ denote the k-th order statistic, and set $U := G(Y_{(k)})$. Then:* (a) $U \sim \mathrm{Beta}(k, n+1-k)$*;* (b) *for every $q \in (0,1)$, the Beta–binomial identity holds:*

$$\mathbb{P}\big(U \le q\big) \;=\; \mathbb{P}\big(\mathrm{Bin}(n,q) \ge k\big);$$

(c) *if $k > nq$, then*

$$\mathbb{P}\big(U \le q\big) \;\le\; \exp\Big(-\frac{2\,(k - nq)^2}{n}\Big).$$

Applicability requires continuity of G (so that the probability integral transform is exact) and the i.i.d. assumption; exchangeability of the sample suffices for (a)–(b).

Proof. (a) Under continuity, $G(Y_i)$ are i.i.d. $\mathrm{Unif}[0,1]$ and G is nondecreasing, so U is the k-th uniform order statistic, whose law is $\mathrm{Beta}(k, n+1-k)$. (b) $\{U \le q\} = \{G(Y_{(k)}) \le q\} = \{\text{at least } k \text{ of the } G(Y_i) \text{ are } \le q\}$, and the indicators $\mathbb{I}\{G(Y_i) \le q\}$ are i.i.d. Bernoulli(q). (c) $\mathrm{Bin}(n,q)$ is a sum of n i.i.d. Bernoulli variables (range 1), so Hoeffding's one-sided bound (Lemma 2 with $R=1$) applied at mean nq gives $\mathbb{P}(\mathrm{Bin}(n,q) \ge k) \le \exp\big(-2(k-nq)^2/n\big)$ whenever $k > nq$. □

Lemma 5 (Consequences of sub-Gaussianity)**.** *Let ϵ be a centered random variable with $\mathbb{E}\big[e^{t\epsilon}\big] \le e^{t^2\sigma^2/2}$ for all $t \in \mathbb{R}$ (variance proxy σ^2). Then:*

(i) $\mathbb{E}[\epsilon^2] \le \sigma^2$;

(ii) $\mathbb{E}\big[e^{2|\epsilon|}\big] \le 2\, e^{2\sigma^2}$;

(iii) $\mathbb{E}\big[\,\big|e^{-\epsilon}-1\big|\,\big] \le \sqrt{2}\,\sigma\, e^{\sigma^2}$.

Proof. (i) Both sides of the mgf bound are finite and smooth near $t=0$; expanding, $1 + \frac{t^2}{2}\mathbb{E}[\epsilon^2] + o(t^2) = \mathbb{E}[e^{t\epsilon}] \le e^{t^2\sigma^2/2} = 1 + \frac{t^2}{2}\sigma^2 + o(t^2)$ (the first-order term vanishes since $\mathbb{E}\epsilon = 0$), and comparing the second-order coefficients as $t \to 0$ gives $\mathbb{E}[\epsilon^2] \le \sigma^2$. (ii) $e^{2|\epsilon|} \le e^{2\epsilon} + e^{-2\epsilon}$ pointwise; taking expectations and applying the mgf bound at $t = \pm 2$ gives $\mathbb{E}e^{2|\epsilon|} \le 2e^{4\sigma^2/2} = 2e^{2\sigma^2}$. (iii) For all $a \in \mathbb{R}$, $|e^a - 1| = \big|\int_0^a e^s\, ds\big| \le |a|\, e^{|a|}$. Applying this with $a = -\epsilon$ and then Cauchy–Schwarz,

$$\mathbb{E}\big|e^{-\epsilon}-1\big| \;\le\; \mathbb{E}\big[|\epsilon|\, e^{|\epsilon|}\big] \;\le\; \sqrt{\mathbb{E}[\epsilon^2]}\,\sqrt{\mathbb{E}\big[e^{2|\epsilon|}\big]} \;\le\; \sigma\,\sqrt{2\,e^{2\sigma^2}} \;=\; \sqrt{2}\,\sigma\, e^{\sigma^2},$$

using (i) and (ii) in the last inequality. □

Remark 1 (Convention on variance proxies). Assumption 3 equips each stratum with a sub-Gaussian proxy σ_b for the centered log scale-ratio; in general the proxy dominates the standard deviation, and the certificate of Theorem 4 identifies the two (moment subtraction recovers the standard deviation). Lemma 5 is stated for an arbitrary proxy; the substitution $\sigma_b = \sqrt{\mathrm{Var}(\lambda \mid b)}$ is the substantive restriction flagged in the main text.

B Deferred Assumptions

The main text states the two structural Assumptions 1 and 2; the remaining technical assumptions are collected here.

Assumption 3 (D: within-stratum sub-Gaussian dispersion)**.** Within each stratum S_b, the centered log scale-ratio $\epsilon_b(X) := \lambda(X) - \bar{\lambda}_b$, with $\bar{\lambda}_b := \mathbb{E}[\lambda(X) \mid X \in S_b]$, is σ_b-sub-Gaussian: $\mathbb{E}[e^{t\epsilon_b}] \le e^{t^2\sigma_b^2/2}$ for all $t \in \mathbb{R}$. The proxy satisfies $\sigma_b \ge \sqrt{\mathrm{Var}(\lambda \mid b)}$ in general; a bounded variant $|\epsilon_b| \le B_b$ replaces it at the cost of a term $e^{B_b} - 1$.

Assumption 4 (B: bounded log-variables)**.** $|W| \le M$ almost surely, where $W := \log(Z/c)$, and $|\lambda(X)| \le M_\lambda$ almost surely, for known constants $M, M_\lambda < \infty$.

Assumption 5 (Q: quantile non-degeneracy)**.** There exist $f_{\min} > 0$ and $r > 0$ such that $f_Z \ge f_{\min}$ on $[u-r,\, u+r]$; equivalently, F_Z^{-1} is $(1/f_{\min})$-Lipschitz near $1-\alpha$.

Assumption 6 (A: Stage I pointwise approximation)**.** The signed Stage-I error $\eta(x) := \log q(x) - \log\big(c\,\sigma_{\mathrm{sim}}(x)\big)$ satisfies $\big|\eta(x)\big| \le \bar{\eta}(x)$ for a known bound $\bar{\eta} : \mathcal{X} \to \mathbb{R}_+$, with $c > 0$ the proportionality constant of the Stage-I objective (Section 3.2, Appendix C); the exact case $\bar{\eta} \equiv 0$, i.e. $q = c\,\sigma_{\mathrm{sim}}$, holds for the population minimizer of equation 14 under Assumption 1.

C Method Details: Stage-I Objective, Simulator Constants, and Instantiations

Stage-I objective. Stage I trains q on simulator residuals $R = |Y - h(X)|$ by a pinball loss at a fixed level $\tau_{\mathrm{tr}} \in (0,1)$ plus an efficiency penalty:

$$q \;\in\; \arg\min_{q \in \mathcal{Q}} \; \mathbb{E}_{\mathbb{P}}\big[\,\rho_{\tau_{\mathrm{tr}}}\big(R - q(X)\big)\,\big] \;+\; \gamma\, \mathbb{E}_{\mathbb{P}}\big[\,\phi\big(q(X)\big)\,\big], \tag{14}$$

where $\rho_{\tau_{\mathrm{tr}}}(t) = t\,(\tau_{\mathrm{tr}} - \mathbf{1}\{t < 0\})$, $\phi \ge 0$ penalizes inflated scales (e.g. $\phi(q) = q$), and $\gamma \ge 0$ trades off the two. The efficiency term restores the coverage–width trade-off a pure quantile fit ignores.

Simulator constants. Under the multiplicative error model (Assumption 1), $R = \sigma_{\mathrm{sim}}(X) \cdot Z$ with $Z \perp X$ under $\mathbb{P}$, so the minimizer of equation 14 still factors as $q^*(x) = c\,\sigma_{\mathrm{sim}}(x)$: at fixed x, positive homogeneity of $\rho_{\tau_{\mathrm{tr}}}$ reduces equation 14 to the scalar problem $\min_{\tilde{c}>0} \mathbb{E}\big[\rho_{\tau_{\mathrm{tr}}}(Z-\tilde{c})\big] + \gamma\,\tilde{c}$ in $\tilde{c} = q(x)/\sigma_{\mathrm{sim}}(x)$, whose first-order condition for $\phi(q) = q$ is $F_Z(c) = \tau_{\mathrm{tr}} - \gamma$, i.e. $c := F_Z^{-1}(\tau_{\mathrm{tr}} - \gamma)$ (the pure-pinball case $\gamma = 0$ recovers $c = F_Z^{-1}(\tau_{\mathrm{tr}})$); for a general penalty ϕ, c is defined as the scalar this minimization returns. Hence one trained function serves all levels: $V = R/q(X) = Z/c$ under $\mathbb{P}$, so $s = u/c$ gives simulator coverage exactly $1-\alpha$, where $u := F_Z^{-1}(1-\alpha)$. The certificate additionally uses the moments $m_1 := \mathbb{E}[W]$ and $m_2 := \mathrm{Var}(W)$ of $W := \log(Z/c)$; all of (u, c, m_1, m_2) are computable from the simulator to arbitrary precision, analytically or by Monte Carlo. The zero-shot threshold is $s_P := u/c$.

Instantiations. SUIT applies wherever a sampler stands in for the deployment environment and residuals approximately factorize into scale and shape (Section 5). *(i) Urn/occupancy family.* In unseen-species problems (distinct values, species richness), the simulator is an urn model with a prescribed tail law Good & Toulmin (1956), h is a plug-in estimator of the unseen count, σ_{sim} the log-error scale, and Z the occupancy-limit shape of the standardized error. The NDV experiments of Section 5 implement this mapping with strata on the estimated tail index; Appendix D develops the scenario in detail. *(ii) Simulation-based inference (SBI) family.* SBI trains neural posterior estimators on unlimited parameter–observation pairs from an intractable-likelihood generative model Tejero-Cantero et al. (2020); Lueckmann et al. (2021); here h is an amortized posterior point estimate, σ_{sim} the posterior dispersion given the observation, and Z the standardized posterior-error shape.

Why this decomposition? No global raw-score threshold, even optimally calibrated, repairs shape heterogeneity, and end-to-end training forfeits amortization (Proposition 1). Dispersion, not level, governs coverage loss (Theorem 2), and stratification localizes it at a quantifiable price (Theorems 6, 7). Real labels are the bottleneck and zero-shot intervals are tightest when the simulator is right; the certificate of Theorem 4 decides when repair pays.

D A Motivating Scenario: Distinct-Value Estimation from Samples

This appendix develops the distinct-value (NDV) scenario that anchors the NDV experiments of Section 5, as a concrete instance of the urn/occupancy instantiation of Appendix C. It spells out what the simulator is in this domain, why learning inside it is attractive, and where the sim-to-real gap enters.

The estimation problem. A database column with N rows contains D distinct values with unknown multiplicities. Drawing a without-replacement sample of $n = rN$ rows at a small rate r (below one percent in practice) yields the sample profile $f = (f_1, f_2, \dots)$, where f_j counts values seen exactly j times; the task is to turn the profile into an interval estimate of D. NDV estimates feed query optimizers, where they drive memory-sensitive operator choices: in the PostgreSQL study underlying our benchmark, a 3% NDV point error pushed a hash-aggregation plan past its memory budget and inflated execution cost by 28%, so a calibrated interval—not a point estimate—is the actionable object. The label cost is the bottleneck: one labeled calibration point means one exact `COUNT(DISTINCT)` pass over a column of up to 3×10^8 rows, so each new column, sampling rate, or coverage level can afford only a handful of real labels.

The simulator. Random sampling from a column is an urn scheme: distinct values are boxes, their multiplicities box probabilities, and the sample profile is an occupancy vector whose asymptotics are governed by the tail of the frequency distribution Good & Toulmin (1956); Karlin (1967); Ben-Hamou et al. (2017). Under a regularly varying tail with index $\alpha \in [0, 1]$—$\alpha = 0$ geometric-like light tails, $\alpha \to 1$ singleton-dominated columns, power-law (Zipfian) columns in between—the conditional error law of any profile-based estimator depends on the column only through coarse, profile-estimable quantities: the tail index (consistently estimated by the singleton ratio $\hat{\alpha} = f_1/d$), the sample size, and a slowly varying factor. A stochastic simulator is therefore available off the shelf: generate synthetic frequency vectors spanning the tail regimes, following the recipe of Wu & Yang (2019); draw sample profiles from them; and read off unlimited labeled triples $(\text{profile}, \text{true } D, \text{estimator error})$ without touching a real database.

Why the uncertainty is heteroscedastic, and why learn the shape in the simulator. The estimation error is strongly input-dependent: occupancy theory gives $\mathrm{Var}(\hat{D}/D) = \Theta\big(N^{\alpha-2}\,\ell(N)\big)$ for $\alpha \in (0, 1)$ Karlin (1967); Ben-Hamou et al. (2017), so between a light-tailed column ($\alpha = 0.2$) and a heavy-tailed one ($\alpha = 0.8$) at $N = 10^8$ the error variance differs by a factor of $N^{0.6} \approx 10^5$, before any variation in N or r. No global interval half-width can be simultaneously valid and informative across such a range, and the error scale is a complicated function of the full profile rather than of any single statistic—precisely the shape–scalar regime of Section 3. Learning the scale on real

data would spend the scarce resource (exact `COUNT(DISTINCT)` labels) on the one part of the problem the simulator already enumerates exhaustively; the sim-to-real gap, by contrast, enters through how closely real columns match the assumed tail family—a low-dimensional residual that the scalar correction of Stage II is designed to absorb.

The pipeline used in our experiments. The NDV experiments instantiate SUIT with: h a learned profile-to-NDV estimator trained on urn simulations; profile features $\phi(x) = (\hat{\alpha},\ \log K_{n,1},\ \log n,\ \log K_{n,\geq 2})$; a $\sim$128K-parameter MLP for q, trained once with the pinball-plus-efficiency objective equation 14 in $\approx$16 minutes and answering in $\approx$1.5 ms per query; and $K = 3$ strata on $\hat{\alpha}$, matching the light/medium/heavy regimes of the variance scaling law. Three observations from this pipeline recur throughout the paper. *(i)* Stratification matters: across the seven benchmark databases, per-stratum scalars yield intervals 1.3–3.7$\times$ narrower than a single global scalar at equal coverage. *(ii)* Zero-shot is nearly safe here: with no real calibration at all, the simulator-internal threshold retains $\approx$0.90 marginal coverage at near-oracle width ($\approx$1.04$\times$)—the simulator is faithful enough that the residual gap is small, and the certificate of Theorem 4 is the instrument that verifies this on a new column. *(iii)* Heavy tails are where methods separate: in the heavy-tailed stratum, global-threshold baselines collapse to roughly half the nominal coverage while the learned shape maintains it (Section 5).

E THE SUIT ALGORITHM

Algorithm 1 SUIT: simulator-trained conformal prediction

Require: Simulator $\mathbb{P}$; predictor h; strata $\{S_b\}_{b=1}^{K}$; level $1-\alpha$; tolerance $\tau \geq 0$
Ensure: Intervals $C(x) = [\,h(x) \pm s_{b(x)}\, q(x)\,]$
Stage I: amortized scale learning (simulator only).
Train q via equation 14 on simulator samples; record c with $q \approx c\,\sigma_{\mathrm{sim}}$.
Compute F_Z, $u = F_Z^{-1}(1-\alpha)$, moments m_1, m_2 of $\log(Z/c)$; set $s_P \leftarrow u/c$.
Stage II: scalar calibration (optional; $n_{\mathrm{cal}} \geq 0$).
$V_i \leftarrow |Y_i - h(X_i)|/q(X_i)$; per stratum b, $\hat{s}_b \leftarrow k_b$-th order statistic of $\{V_i : X_i \in S_b\}$, $k_b = \lceil (n_b+1)(1-\alpha) \rceil$; if $n_{\mathrm{cal}} = 0$, deploy s_P (zero-shot).
Deployment: certificate-based selective calibration.
for each stratum b, periodically **do**
  Update Cert_b (Theorem 4).
  if shape audit rejects **then**
    Recalibrate stratum b; warn of simulator mismatch.
  else if $\mathrm{Cert}_b \geq 1-\alpha-\tau$ **then**
    Keep zero-shot: $s_b \leftarrow s_P$.
  else
    Scalar recalibration: $s_b \leftarrow \hat{s}_b$ from stratum-b points.
  end if
end for

F PROOF OF THEOREM 1 AND COROLLARY 1

Proof of Theorem 1. We use Assumptions 1 and 2 and the exact Stage-I relation $q = c\,\sigma_{\mathrm{sim}}$. By the key identity equation 4, for every $x \in \mathcal{X}$,

$$\mathbb{Q}\big(V \leq s_P \mid X = x\big) \;=\; F_Z\big(c\,s_P\,e^{-\lambda(x)}\big) \;=\; F_Z\big(u\,e^{-\lambda(x)}\big),$$

since $c\,s_P = u$. Taking expectations over $X \sim \mathbb{Q}^X$ and inserting $F_Z(u) = 1-\alpha$,

$$\mathrm{Cov}_{\mathrm{real}}(s_P) \;=\; \mathbb{E}_X\Big[F_Z\big(u\,e^{-\lambda(X)}\big)\Big] \;\geq\; (1-\alpha) \;-\; \mathbb{E}_X\Big[\,\big|F_Z\big(u\,e^{-\lambda(X)}\big) - F_Z(u)\big|\,\Big].$$

We bound the integrand pointwise in x. Since f_Z exists with $\sup_z f_Z(z) = C_f$ (Assumption 2), the mean-value theorem gives ξ between u and $ue^{-\lambda(x)}$ with

$$F_Z\big(u\,e^{-\lambda(x)}\big) - F_Z(u) \;=\; f_Z(\xi)\,u\,\big(e^{-\lambda(x)} - 1\big), \qquad \text{hence} \qquad \big|F_Z\big(u\,e^{-\lambda(x)}\big) - F_Z(u)\big| \;\leq\; C_f\,u\,\big|e^{-\lambda(x)} - 1\big|.$$

Taking expectations over X yields the first inequality of equation 5. For the second, the mean-value theorem applied to $s \mapsto e^{-s}$ gives $\big|e^{-\lambda(x)} - 1\big| \leq e^{|\lambda(x)|}\,|\lambda(x)| \leq e^{\|\lambda\|_\infty}\,|\lambda(x)|$, and inserting this pointwise bound into the first inequality completes the proof. □

Proof of Corollary 1. Under pure covariate shift the conditional error model is shared, $\sigma_{\mathrm{sim}} = \sigma_{\mathrm{real}}$, so $\lambda(x) = \log\big(\sigma_{\mathrm{real}}(x)/\sigma_{\mathrm{sim}}(x)\big) \equiv 0$ on the union of the two covariate supports. The key identity equation 4 therefore gives, for every x in the real support,

$$\mathbb{Q}\big(V \le s_P \mid X = x\big) \;=\; F_Z(u) \;=\; 1-\alpha,$$

and averaging over $X \sim \mathbb{Q}^X$ yields $\mathrm{Cov}_{\mathrm{real}}(s_P) = 1-\alpha$ exactly, irrespective of $\mathbb{Q}^X$. The same computation under $\mathbb{P}$ (where $V = Z/c$) gives $\mathrm{Cov}_{\mathrm{sim}}(s_P) = F_Z(c\,s_P) = F_Z(u) = 1-\alpha$. For the maximal-discrepancy clause: if $\mathbb{P}^X$ and $\mathbb{Q}^X$ have disjoint supports, then the joint laws $\mathbb{P}^{X,Y}$, $\mathbb{Q}^{X,Y}$ have disjoint supports as well, and choosing A to be the support of $\mathbb{P}^{X,Y}$ gives $\mathrm{TV}\big(\mathbb{P}^{X,Y},\mathbb{Q}^{X,Y}\big) \ge |\mathbb{P}^{X,Y}(A) - \mathbb{Q}^{X,Y}(A)| = 1$; disjoint supports likewise preclude an absolutely-continuous density ratio $w = d\mathbb{Q}/d\mathbb{P}$ on the union, so the weighted-conformal degradation term is undefined or unbounded there, while the pushforward slack of Lemma 1 is exactly zero since the laws of V coincide. □

Remark 2 (Structural extrapolation). The exactness in Corollary 1 requires q, trained on simulator covariates, to remain evaluable and equal to $c\,\sigma_{\mathrm{sim}}$ on the *real* support of X; this structural-extrapolation requirement is stated in the main text and is not verifiable from simulator data alone.

G Proof of Theorem 2 and Corollary 2

We first isolate the finite-sample relaxation of the marginal quantile, with conditions stated explicitly.

Lemma 6 (Finite-sample relaxation of the Stage-II scalar)**.** *Let $V_1,\dots,V_{n_{\mathrm{cal}}}$ be i.i.d. draws from $\mathbb{Q}^V$ with marginal cdf $G(t) = \mathbb{Q}(V \le t)$, assumed continuous (which holds under Assumption 2, since $G(t) = \mathbb{E}_X\big[F_Z\big(c\,t\,e^{-\lambda(X)}\big)\big]$ is a mixture of the continuous, strictly increasing functions $t \mapsto F_Z(ct\,e^{-\lambda(x)})$). Let $\hat{s} = V_{(k)}$ with $k = \lceil (n_{\mathrm{cal}}+1)(1-\alpha)\rceil$; if $k = n_{\mathrm{cal}}+1$ set $\hat{s} = +\infty$ and all statements below hold trivially. Then, with probability at least $1-\beta$ over the calibration set,*

$$G(\hat{s}) \;\ge\; 1-\alpha \;-\; \epsilon_n, \qquad \epsilon_n \;:=\; \sqrt{\frac{\ln(2/\beta)}{2\,(n_{\mathrm{cal}}+1)}},$$

and, with probability at least $1-\beta$ (adjusting the constant inside the logarithm), the two-sided statement $|G(\hat{s}) - (1-\alpha)| \le \epsilon_n + 1/(n_{\mathrm{cal}}+1)$ holds. In particular $\epsilon_n = \sqrt{\ln(2/\beta)/(2n_{\mathrm{cal}})} + O(1/n_{\mathrm{cal}})$, matching the convention used in the main text.

Proof. Continuity of G licenses Lemma 4: $U := G(\hat{s}) \sim \mathrm{Beta}(k, n_{\mathrm{cal}}+1-k)$, where $n := n_{\mathrm{cal}}$. Set $c := k/(n+1) \ge 1-\alpha$ (the ceiling). For $0 < t \le c$, apply Lemma 4(b)–(c) with $q = c - t$: the binomial mean is $nq = nc - nt$ and the mean-to-threshold gap is

$$k - nq \;=\; (n+1)c - nc + nt \;=\; c + nt \;\ge\; (n+1)\,t,$$

since $c \ge t$. Hence

$$\mathbb{P}\big(U \le c - t\big) \;\le\; \exp\Big(-\frac{2\,(c+nt)^2}{n}\Big) \;\le\; \exp\big(-2\,(n+1)\,t^2\big),$$

because $(c+nt)^2 \ge (n+1)^2t^2 \ge n(n+1)t^2$. (If $t > c$ then $\mathbb{P}(U \le c-t) = 0$ and the same display is trivially valid.) Taking $t = \epsilon_n$ makes the right side $\beta/2 \le \beta$, and on the complementary event $G(\hat{s}) > c - \epsilon_n \ge 1 - \alpha - \epsilon_n$; this proves the lower-tail statement with probability at least $1 - \beta/2 \ge 1-\beta$.

For the two-sided statement, apply Lemma 4 to the upper tail: $\mathbb{P}(U \ge c+t) = \mathbb{P}\big(\mathrm{Bin}(n, c+t) \le k-1\big)$ for $c + t < 1$, whose mean-to-threshold gap is $n(c+t) - (k-1) = nt + 1 - c \ge nt$ (as $c \le 1$), giving $\mathbb{P}(U \ge c+t) \le e^{-2(nt+1-c)^2/n} \le e^{-2nt^2}$. Setting $t = \sqrt{\ln(4/\beta)/(2n)}$ bounds each tail by $\beta/2$; combining with $c - (1-\alpha) \in [0, 1/(n+1))$ yields $|G(\hat{s}) - (1-\alpha)| \le \epsilon_n' + 1/(n+1)$ with probability at least $1-\beta$, where ϵ_n' differs from ϵ_n only by constants inside the logarithm and the $(n+1)$ versus n denominator. Finally, $\sqrt{1/(n+1)} = \sqrt{1/n}\,\big(1 + O(1/n)\big)$, so $\epsilon_n = \sqrt{\ln(2/\beta)/(2n)} + O(1/n)$ as claimed. □

Proof of Theorem 2. We use Assumptions 1 and 2 and $q = c\,\sigma_{\mathrm{sim}}$; Assumption 5 is invoked only in Step 3. Write $n := n_{\mathrm{cal}}$ and condition on the calibration set throughout: $\hat{s}$ is then a fixed scalar.

Step 1: exact representations. Since $\lambda(x) = \bar{\lambda} + \epsilon(x)$ and $d(x) = e^{-\epsilon(x)}$, the key identity equation 4 gives, for every x,

$$\mathrm{Cov}_{\mathrm{real}}(\hat{s} \mid X = x) \;=\; F_Z\big(c\,\hat{s}\,e^{-\bar{\lambda}}\,e^{-\epsilon(x)}\big) \;=\; F_Z\big(\tilde{t}\,d(x)\big), \qquad G(\hat{s}) \;=\; \mathbb{E}_{X'}\Big[F_Z\big(\tilde{t}\,d(X')\big)\Big],$$

the second display by averaging the first over $X' \sim \mathbb{Q}^X$.

Step 2: dispersion decomposition. Adding and subtracting $F_Z(\tilde{t})$ and $\mathbb{E}_{X'}[F_Z(\tilde{t}\, d(X'))] - F_Z(\tilde{t})$,

$$\mathrm{Cov}_{\mathrm{real}}(\hat{s} \mid x) \;=\; G(\hat{s}) \;+\; \big[F_Z(\tilde{t}\, d(x)) - F_Z(\tilde{t})\big] \;-\; \mathbb{E}_{X'}\big[F_Z(\tilde{t}\, d(X')) - F_Z(\tilde{t})\big].$$

By the mean-value theorem and $\sup_z f_Z(z) = C_f$ (Assumption 2), $\big|F_Z(a) - F_Z(b)\big| \le C_f\, |a-b|$, applied pointwise to each bracket:

$$\big|F_Z(\tilde{t}\, d(x)) - F_Z(\tilde{t})\big| \;\le\; C_f\, \tilde{t}\, \big|d(x) - 1\big| \;=\; C_f\, \tilde{t}\, \big|e^{-\epsilon(x)} - 1\big|,$$

and identically for the expectation term, where taking absolute values inside the expectation is legitimate by Jensen's inequality for $|\cdot|$. Therefore

$$\mathrm{Cov}_{\mathrm{real}}(\hat{s} \mid x) \;\ge\; G(\hat{s}) \;-\; C_f\, \tilde{t}\, \Big(\big|e^{-\epsilon(x)} - 1\big| + \mathbb{E}_{X'}\big|e^{-\epsilon(X')} - 1\big|\Big).$$

Lemma 6 (whose continuity requirement holds by Assumption 2) gives $G(\hat{s}) \ge 1 - \alpha - \epsilon_n$ with probability at least $1 - \beta$ over the calibration set, and substituting yields equation 7. The event is determined by the calibration set alone and the bound holds simultaneously for all x, since Steps 1–2 are pointwise. The symmetric upper bound follows from the same display with the roles of the two brackets reversed:

$$\mathrm{Cov}_{\mathrm{real}}(\hat{s} \mid x) \;\le\; 1 - \alpha \;+\; C_f\, \tilde{t}\, \Big(\big|e^{-\epsilon(x)} - 1\big| + \mathbb{E}_{X'}\big|e^{-\epsilon(X')} - 1\big|\Big) \;+\; \epsilon_n,$$

using the two-sided part of Lemma 6 (constant inside the logarithm adjusted).

Step 3: order of $\tilde{t}$ under Assumption 5. Define $H(t) := \mathbb{E}_{X'}\big[F_Z\big(t\, d(X')\big)\big]$ and let s^* denote a population marginal quantile, $G(s^*) = 1 - \alpha$ (existence: G is continuous with $G(0) = 0$, $G(\infty) = 1$); set $t^* := c\, s^*\, e^{-\bar{\lambda}}$, so that $H(t^*) = G(s^*) = 1 - \alpha = F_Z(u) = H(u) + \big[F_Z(u) - H(u)\big]$.

(i) Function-value proximity. By the Lipschitz bound of Step 2 applied inside the expectation,

$$\big|H(u) - F_Z(u)\big| \;\le\; C_f\, u\, \mathbb{E}_{X'}\big|d(X') - 1\big| \;=:\; \delta_1.$$

(ii) Local strong monotonicity. Write $E := \|\epsilon\|_\infty$, so $d(X') \in [e^{-E}, e^{E}]$ almost surely. On any segment J joining u and t^* with the property that $t\, d(X') \in [u - r, u + r]$ for all $t \in J$ almost surely,

$$H'(t) \;=\; \mathbb{E}_{X'}\Big[f_Z\big(t\, d(X')\big)\, d(X')\Big] \;\ge\; f_{\min}\, e^{-E} \;>\; 0,$$

by Assumption 5. To see that the relevant segment has this property, first sandwich H: since F_Z is nondecreasing and $d(X') \in [e^{-E}, e^{E}]$, $F_Z(t\, e^{-E}) \le H(t) \le F_Z(t\, e^{E})$ for all $t \ge 0$; evaluating at t^* gives $F_Z(t^* e^{-E}) \le F_Z(u) \le F_Z(t^* e^{E})$, hence $u\, e^{-E} \le t^* \le u\, e^{E}$, and every t between u and t^* satisfies $t\, d(X') \in [u\, e^{-2E},\, u\, e^{2E}]$. The derivative bound is therefore valid on the whole segment whenever

$$\big[u\, e^{-2E},\, u\, e^{2E}\big] \;\subseteq\; [u - r,\, u + r], \tag{15}$$

which holds in the small-dispersion regime to which the order statement pertains, in particular whenever $u\,(e^{2E} - 1) \le r$ (a fortiori for $E \le \frac{1}{2}\ln(1 + r/u)$), and this is the quantification of "a neighborhood of u" in the theorem statement. Granting equation 15, the mean-value theorem applied to H on the segment joining u and t^* gives

$$\big|H(t^*) - H(u)\big| \;\ge\; f_{\min}\, e^{-E}\, \big|t^* - u\big|,$$

while $H(t^*) = 1 - \alpha = F_Z(u)$ exactly, so $\big|H(t^*) - H(u)\big| = \big|F_Z(u) - H(u)\big| \le \delta_1$ by step (i). Combining,

$$|t^* - u| \;\le\; \delta_1\, e^{E} / f_{\min} \;<\; r\, e^{-E},$$

where the last inequality uses the condition $\delta_1 < f_{\min}\, r\, e^{-E}$ (which gives $\delta_1 e^{E} / f_{\min} < r$); this a posteriori places t^* strictly inside the radius-r neighborhood of u, consistent with the locality regime of equation 15. Substituting δ_1 and $|d(X') - 1| = |e^{-\epsilon(X')} - 1| \le e^{E}\, |\epsilon(X')|$ (mean-value theorem on the exponential),

$$|t^* - u| \;\le\; \frac{C_f}{f_{\min}}\, u\, e^{E}\, \mathbb{E}\big|d(X') - 1\big| \;\le\; \frac{C_f}{f_{\min}}\, u\, e^{2E}\, \mathbb{E}|\epsilon|.$$

(iii) Transporting the estimation error. On the probability-$(1 - \beta)$ event of Lemma 6 (two-sided form), $|G(\hat{s}) - G(s^*)| \le \epsilon_n + 1/(n+1)$. Now $G(s) = H(c\, s\, e^{-\bar{\lambda}})$, so $G'(s) = c\, e^{-\bar{\lambda}}\, H'(c\, s\, e^{-\bar{\lambda}}) \ge c\, e^{-\bar{\lambda}}\, f_{\min}\, e^{-E}$ on the segment joining s^* and $\hat{s}$ whenever that segment maps into the region of step (ii); the same locality condition equation 15

suffices (the segment lies within $[ue^{-E}, ue^{E}] \cdot (ce^{-\bar{\lambda}})^{-1}$ after rescaling, and the two-sided control keeps $\hat{s}$ within the region where this applies, up to the same small-dispersion qualification). Hence

$$|\tilde{t} - t^*| \;=\; c\,e^{-\bar{\lambda}}\,|\hat{s} - s^*| \;\le\; \frac{c\,e^{-\bar{\lambda}}\,\big(\epsilon_n + 1/(n+1)\big)}{c\,e^{-\bar{\lambda}}\,f_{\min}\,e^{-E}} \;=\; \frac{\epsilon_n\,e^{E}}{f_{\min}}\,\big(1 + O(1/n)\big).$$

Combining (ii) and (iii) via $|\tilde{t} - u| \le |\tilde{t} - t^*| + |t^* - u|$,

$$\tilde{t} \;=\; u\,\Big(1 + O\big(\mathbb{E}|\epsilon| + \epsilon_n\big)\Big),$$

where the dispersion term carries implicit constant $(C_f/f_{\min})\,e^{O(\|\epsilon\|_\infty)}$ and the ϵ_n term carries relative constant $e^{\|\epsilon\|_\infty}/(u\,f_{\min})$, as stated. In particular $C_f\,\tilde{t} = C_f\,u\,(1 + o(1))$ whenever $\mathbb{E}|\epsilon| \to 0$ and $\epsilon_n \to 0$. □

Remark 3 (Assumption 5 is not needed for the bounds). Neither the main inequality equation 7 nor Corollary 2 invokes Assumption 5: equation 7 holds with $\tilde{t} = c\,\hat{s}\,e^{-\bar{\lambda}}$ itself as the constant, by Steps 1–2 alone. Assumption 5 enters only Step 3, the asymptotic reading $\tilde{t} \approx u$ with non-exploding constants, which is the direction in which function-value proximity $\big(H(u) \approx F_Z(u)\big)$ can be converted into argument proximity $\big(t^* \approx u\big)$: the inverse-cdf Lipschitz direction. The locality condition equation 15 is automatic in the small-dispersion regime and is stated explicitly above; it does not alter any constant in the theorem.

Proof of Corollary 2. Pure global scale shift means $\sigma_{\text{real}} = e^{\bar{\lambda}}\sigma_{\text{sim}}$, i.e. $\epsilon(x) = \lambda(x) - \bar{\lambda} \equiv 0$; no restriction on the magnitude of $\bar{\lambda}$ is imposed. Then both dispersion terms in equation 7 vanish identically: $\big|e^{-\epsilon(x)} - 1\big| = 0$ for every x and likewise for the expectation over X'. Theorem 2 (which does not require Assumption 5 for the inequality itself) therefore gives, with probability at least $1 - \beta$ over the calibration set, simultaneously for every x,

$$\mathrm{Cov}_{\text{real}}(\hat{s} \mid X = x) \;\ge\; 1 - \alpha - \epsilon_n, \qquad \epsilon_n = \sqrt{\tfrac{\ln(2/\beta)}{2(n_{\text{cal}}+1)}} = O\Big(\sqrt{\log(1/\beta)/n_{\text{cal}}}\Big).$$

Solving $\epsilon_n \le \delta$ gives the sufficient sample size $n_{\text{cal}} \ge \ln(2/\beta)/(2\delta^2) - 1 = O\big(\delta^{-2}\log(1/\beta)\big)$, with constants independent of the shift magnitude $|\bar{\lambda}|$. □

H Statement and Proof of Theorem 3

We restate the theorem in full.

Theorem 3 (Shape mismatch hurts only the zero-shot mode)**.** *Retain Assumption 1 and $q = c\,\sigma_{\text{sim}}$, but drop Assumption 2, allowing $\mathcal{L}_{\text{sim}}(Z) \neq \mathcal{L}_{\text{real}}(Z)$.*

(a) Zero-shot bound. *Inequality equation 5 becomes*

$$\mathrm{Cov}_{\text{real}}(s_P) \;\ge\; 1 - \alpha \;-\; C_f\,u\,\mathbb{E}_X\Big[\,\big|e^{-\lambda(X)} - 1\big|\,\Big] \;-\; \mathrm{KS}\big(\mathcal{L}_{\text{sim}}(Z),\,\mathcal{L}_{\text{real}}(Z)\big), \tag{16}$$

where KS *is the Kolmogorov–Smirnov distance* $\sup_z |F_Z^{\text{sim}}(z) - F_Z^{\text{real}}(z)|$*; the same statement with* TV *in place of* KS *holds as a conservative corollary.*

(b) Two-stage bounds are unaffected. *Theorem 2 and Corollary 2 continue to hold without any relation between the two shape laws, requiring only $Z_{\text{real}} \perp X$ under $\mathbb{Q}$.*

Proof of Theorem 3. We retain Assumption 1 in both environments and $q = c\,\sigma_{\text{sim}}$, but drop Assumption 2, writing F_Z^{sim} and F_Z^{real} for the two shape cdf's; the constants $u = (F_Z^{\text{sim}})^{-1}(1 - \alpha)$ and $C_f = \sup_z f_Z^{\text{sim}}(z)$ belong to the simulator shape.

(a) Zero-shot bound. By the key identity equation 4 applied under $\mathbb{Q}$ with the *real* shape,

$$\begin{aligned}\mathrm{Cov}_{\text{real}}(s_P) \;&=\; \mathbb{E}_X\Big[F_Z^{\text{real}}\big(u\,e^{-\lambda(X)}\big)\Big] \\ &\ge\; \mathbb{E}_X\Big[F_Z^{\text{sim}}\big(u\,e^{-\lambda(X)}\big)\Big] \;-\; \sup_z\,\big|F_Z^{\text{real}}(z) - F_Z^{\text{sim}}(z)\big|,\end{aligned}$$

where the second step holds pointwise in x, since the integrands differ by at most the supremum; that supremum is exactly the Kolmogorov–Smirnov distance $\mathrm{KS}\big(\mathcal{L}_{\text{sim}}(Z), \mathcal{L}_{\text{real}}(Z)\big)$. The simulator-shaped expectation is precisely the

quantity analyzed in the proof of Theorem 1: inserting $F_Z^{\mathrm{sim}}(u) = 1-\alpha$ and applying, pointwise in x, the mean-value bound

$$\left|F_Z^{\mathrm{sim}}\big(ue^{-\lambda(x)}\big) - F_Z^{\mathrm{sim}}(u)\right| \;\le\; C_f\, u\,\left|e^{-\lambda(x)} - 1\right|$$

yields

$$\mathrm{Cov}_{\mathrm{real}}(s_P) \;\ge\; 1-\alpha \;-\; C_f\, u\, \mathbb{E}_X\left|e^{-\lambda(X)} - 1\right| \;-\; \mathrm{KS}\big(\mathcal{L}_{\mathrm{sim}}(Z), \mathcal{L}_{\mathrm{real}}(Z)\big),$$

which is equation 16. Since $\mathrm{KS} \le \mathrm{TV}$ for any two laws (the KS distance is a supremum over the subclass of half-line events), the same display with $\mathrm{TV}\big(\mathcal{L}_{\mathrm{sim}}(Z), \mathcal{L}_{\mathrm{real}}(Z)\big)$ in place of KS holds as a conservative corollary.

(b) Two-stage bounds. Stage II uses only real calibration data. The identity equation 4 is valid under $\mathbb{Q}$ with F_Z^{real} in place of F_Z, since it uses only $V = \rho(X)Z/c$ and $Z_{\mathrm{real}} \perp X$ (Assumption 1); no relation between the two shape laws enters. With this substitution, the proof of Theorem 2 goes through line by line: the exact representations of Step 1 hold with F_Z^{real}; the dispersion decomposition of Step 2 uses only the Lipschitz constant of the *real* shape, $C_{f,R} := \sup_z f_Z^{\mathrm{real}}(z)$; and the finite-sample relaxation of Lemma 6 is distribution-free: the marginal coverage of $\hat{s}$ rests on exchangeability of the real calibration sample, which makes no shape assumption at all. Theorem 2 and Corollary 2 therefore hold verbatim without Assumption 2, with the real-shape constants $u_R = (F_Z^{\mathrm{real}})^{-1}(1-\alpha)$ and $C_{f,R}$ replacing u and C_f (and the reading $\tilde{t} \approx u$ adapted accordingly, requiring the real-shape analogue of Assumption 5). □

Remark 4 (Efficiency versus coverage under shape mismatch). Part (b) concerns coverage only. Shape mismatch still degrades *efficiency*: q is trained on simulator shape quantiles, so if the real shape is more dispersed, $\hat{s}$ inflates to compensate and intervals widen while calibrated coverage is preserved. Without Assumption 2 the constants $u_R, C_{f,R}$ are not identifiable from the simulator, which is why we state all subsequent results under Assumption 2 with simulator-computable constants; the audit of Proposition 4 is the instrument that detects the mismatch regime.

I Statements and Proofs of Propositions 1 and 2

We restate both propositions in full.

Proposition 1 (Impossibility of raw-score scalar thresholds)**.** *Consider any fixed global threshold $t > 0$ applied to the raw score, $C(x) = [\,h(x) \pm t\,]$, and allow t to be chosen optimally with unlimited real calibration data. Under Assumption 1, suppose $\sigma_{\mathrm{real}}(X)$ takes two values $\sigma_{\mathrm{lo}} < \sigma_{\mathrm{hi}}$ with $p := \mathbb{Q}\big(\sigma(X) = \sigma_{\mathrm{hi}}\big) \in (0,1)$, and that $f_Z \ge f_{\min} > 0$ on the relevant interval. Then the marginally optimal threshold t^*, defined by $\mathbb{E}\big[F_Z\big(t^*/\sigma(X)\big)\big] = 1-\alpha$, satisfies*

(a) Two-point coverage defect.

$$\mathrm{Cov}\big(t^* \,\big|\, \sigma = \sigma_{\mathrm{hi}}\big) \;\le\; 1-\alpha \;-\; (1-p)\, f_{\min}\, t^* \left(\tfrac{1}{\sigma_{\mathrm{lo}}} - \tfrac{1}{\sigma_{\mathrm{hi}}}\right); \tag{17}$$

(b) Unrepairability. *The defect in equation 17 is first-order in the scale span $\sigma_{\mathrm{hi}}/\sigma_{\mathrm{lo}}$ and independent of the calibration sample size; uniform pointwise conditional coverage would require $t/\sigma(x)$ to be constant in x, which no scalar t achieves unless σ itself is constant.*

In the urn instantiation of Section 5, the error scale spans orders of magnitude across tail regimes—the variance scaling law of Appendix D gives a factor of $N^{0.6} \approx 10^5$ between light- and heavy-tailed columns at $N = 10^8$—and global-score baselines collapse to roughly half the nominal coverage in the heavy-scale stratum: the defect of equation 17 is not a worst-case curiosity but the realized behavior of raw-score calibration in the paper's central application family.

Proposition 2 (Separation from weighted conformal prediction)**.** *Recall that weighted split conformal prediction Tibshirani et al. (2019) achieves exact marginal coverage under covariate shift when the density ratio $w = d\mathbb{Q}/d\mathbb{P}$ is known, and degrades to $1-\alpha - \frac{1}{2}\mathbb{E}|\hat{w} - w|$ with an estimated ratio $\hat{w}$. The two-stage structure separates from this route in three ways.*

(a) Exactness where the ratio bound fails. *Under pure covariate shift with disjoint supports, w is unbounded or undefined and the term $\frac{1}{2}\mathbb{E}|\hat{w} - w|$ is vacuous, while* SUIT *attains exact coverage, already at $n_{\mathrm{cal}} = 0$ (Corollary 1).*

(b) Sample complexity. *Stage II estimates one scalar, a one-dimensional quantile of V, at the distribution-free DKW rate $\epsilon_n = O\big(\sqrt{\ln(1/\beta)/n_{\mathrm{cal}}}\big)$, with constants independent of the shift magnitude. Model-agnostic estimation of $\hat{w}$ is instead a $(d{+}1)$-dimensional function-estimation problem: even in the most favorable*

parametric case, driving $\mathbb{E}|\hat{w} - w|$ to δ_w requires $n_{\mathrm{cal}} = \Omega(\delta_w^{-2})$ with constants depending on $\|w\|_\infty$, and under a scale shift of magnitude $|\bar{\lambda}|$ the joint ratio

$$w(x, r) \;=\; \frac{\sigma_{\mathrm{sim}}(x)}{\sigma_{\mathrm{real}}(x)} \cdot \frac{f_Z\big(r/\sigma_{\mathrm{real}}(x)\big)}{f_Z\big(r/\sigma_{\mathrm{sim}}(x)\big)}$$

has sup-norm growing with $|\bar{\lambda}|$; nonparametric minimax rates deteriorate further with dimension.

(c) Conditional versus marginal. *Weighted calibration restores marginal coverage, but the deployed threshold remains a global scalar, so its conditional coverage under heteroscedastic shift is as uneven as the raw-score method's (Proposition 1); Theorem 2 provides pointwise conditional control.*

Proof of Proposition 1. (a) Two-point coverage defect. Under Assumption 1, the pointwise conditional coverage of the raw-score interval $[h(x) \pm t]$ is $\mathrm{Cov}(t \mid x) = \mathbb{Q}\big(|Y - h(X)| \le t \,\big|\, X = x\big) = F_Z\big(t/\sigma(x)\big)$, since $R = \sigma(X)Z$ with $Z \perp X$ and F_Z continuous. The marginally optimal threshold t^* is defined by the marginal constraint

$$p\, F_Z\big(t^*/\sigma_{\mathrm{hi}}\big) + (1-p)\, F_Z\big(t^*/\sigma_{\mathrm{lo}}\big) \;=\; 1 - \alpha,$$

which has a solution since the left side is continuous, increasing in t^*, and ranges from 0 to 1. Solving for the heavy-scale term,

$$F_Z\big(t^*/\sigma_{\mathrm{hi}}\big) \;=\; (1-\alpha) \;-\; (1-p)\Big[F_Z\big(t^*/\sigma_{\mathrm{lo}}\big) - F_Z\big(t^*/\sigma_{\mathrm{hi}}\big)\Big].$$

Since $\sigma_{\mathrm{lo}} < \sigma_{\mathrm{hi}}$, we have $t^*/\sigma_{\mathrm{hi}} < t^*/\sigma_{\mathrm{lo}}$, and the mean-value theorem gives $\xi \in \big(t^*/\sigma_{\mathrm{hi}},\, t^*/\sigma_{\mathrm{lo}}\big)$ with

$$F_Z\big(t^*/\sigma_{\mathrm{lo}}\big) - F_Z\big(t^*/\sigma_{\mathrm{hi}}\big) \;=\; f_Z(\xi)\, t^* \left(\frac{1}{\sigma_{\mathrm{lo}}} - \frac{1}{\sigma_{\mathrm{hi}}}\right) \;\ge\; f_{\min}\, t^* \left(\frac{1}{\sigma_{\mathrm{lo}}} - \frac{1}{\sigma_{\mathrm{hi}}}\right),$$

using the hypothesis $f_Z \ge f_{\min} > 0$ on the relevant interval. Substitution yields equation 17:

$$\mathrm{Cov}\big(t^* \,\big|\, \sigma = \sigma_{\mathrm{hi}}\big) \;=\; F_Z\big(t^*/\sigma_{\mathrm{hi}}\big) \;\le\; 1 - \alpha \;-\; (1-p)\, f_{\min}\, t^* \left(\tfrac{1}{\sigma_{\mathrm{lo}}} - \tfrac{1}{\sigma_{\mathrm{hi}}}\right).$$

(b) Unrepairability. The subtracted term in equation 17 is first-order in the scale span: $1/\sigma_{\mathrm{lo}} - 1/\sigma_{\mathrm{hi}} = (\sigma_{\mathrm{hi}} - \sigma_{\mathrm{lo}})/(\sigma_{\mathrm{lo}}\sigma_{\mathrm{hi}})$ grows proportionally with the span at fixed σ_{lo}, and the bound involves no sample-size-dependent quantity; t^* is already the population optimum, so additional calibration data cannot improve it. Finally, uniform pointwise conditional coverage $\mathrm{Cov}(t \mid x) = F_Z\big(t/\sigma(x)\big) \ge 1 - \alpha$ for all x with the same t would require $t/\sigma(x)$ to be constant in x on the support, which a scalar t achieves only if σ itself is constant. □

Proof of Proposition 2. (a) Exactness where the ratio bound fails. This is Corollary 1 set against the weighted-conformal degradation bound $1 - \alpha - \frac{1}{2}\mathbb{E}|\hat{w} - w|$ Tibshirani et al. (2019): under pure covariate shift with disjoint supports, $w = d\mathbb{Q}/d\mathbb{P}$ fails to exist on the union of supports (equivalently $\|w\|_\infty$ is unbounded on any common dominating representation), so the degradation term is undefined or vacuous, while the two-stage zero-shot coverage is exactly $1 - \alpha$ as proved in Corollary 1.

(b) Sample complexity. The Stage-II estimation target is the k-th order statistic of a one-dimensional sample: Lemma 6 (via the Beta–binomial identity, Lemma 4) delivers the distribution-free rate $\epsilon_n = O\big(\sqrt{\ln(1/\beta)/n_{\mathrm{cal}}}\big)$ with constants independent of the shift magnitude, the same rate the DKW inequality gives for one-dimensional cdf estimation. The weighted route instead requires estimating the joint density ratio $w = d\mathbb{Q}/d\mathbb{P}$, a function on the $(d{+}1)$-dimensional space of (x, r); even in the most favorable correctly-specified parametric case, the degradation budget $\frac{1}{2}\mathbb{E}|\hat{w} - w| \le \delta_w$ requires $n_{\mathrm{cal}} = \Omega(\delta_w^{-2})$ by the usual L^1 parametric rate, with constants depending on $\|w\|_\infty$. Under a scale shift $\sigma_{\mathrm{real}} = e^{\bar{\lambda}}\sigma_{\mathrm{sim}}$ with a shared shape, the joint ratio has the explicit form

$$w(x, r) \;=\; \frac{\sigma_{\mathrm{sim}}(x)}{\sigma_{\mathrm{real}}(x)} \cdot \frac{f_Z\big(r/\sigma_{\mathrm{real}}(x)\big)}{f_Z\big(r/\sigma_{\mathrm{sim}}(x)\big)},$$

whose sup-norm grows with $|\bar{\lambda}|$ under standard tail behavior of f_Z (e.g. unimodal shapes with monotone or regularly varying tails, for which the density-ratio factor is bounded away from zero on a region whose probability does not vanish with $|\bar{\lambda}|$); we state this growth as a tail-condition-dependent fact rather than a universal theorem. Nonparametric estimation of w fares strictly worse, with minimax rates deteriorating in the dimension $d{+}1$. Hence the two-stage construction reduces the statistical difficulty of the sim-to-real correction from "$(d{+}1)$-dimensional function estimation with constants growing in the shift magnitude" to "one-dimensional quantile estimation with shift-free constants."

(c) Conditional versus marginal. Weighted calibration restores *marginal* coverage $1-\alpha$, but the deployed interval remains $[h(x) \pm t]$ with a single global scalar t (weighting changes the threshold's value, not its functional form). Its conditional coverage under heteroscedastic shift is therefore governed by Proposition 1: the two-point analysis applies verbatim to the weighted-optimal t^*, which is marginally optimal under $\mathbb{Q}$ by construction. Theorem 2, in contrast, controls coverage pointwise in x. □

Remark 5 (Honest boundary of the separation). If the analyst knows the true parametric model (Assumption 1 together with the functional form of ρ), then w is estimable at a parametric rate and the sample-complexity separation of part (b) closes. The separation is therefore claimed between model-free one-dimensional estimation and model-agnostic density-ratio estimation: the two-stage method attains, without model knowledge, the accuracy weighted conformal reaches only with parametric knowledge, while adding pointwise conditional coverage and a zero-shot operating point.

J Statement and Proof of Theorem 4

We restate the theorem in full, including the computable certificate equation 19; the main text presents only the boxed guarantee $\mathrm{Cov}_b(s_P) \geq \mathrm{Cert}_b$.

Theorem 4 (Moment identification and the stratified certificate)**.** *Under Assumptions 1, 2, 3, and 4, with $q = c\,\sigma_{\mathrm{sim}}$:*

(a) Moment identification. *For each stratum b,*

$$\mathbb{E}[\log V \mid b] \;=\; \bar{\lambda}_b + m_1, \qquad \mathrm{Var}[\log V \mid b] \;=\; \sigma_b^2 + m_2,$$

so the first two moments of the scale-ratio distribution are identified in closed form by moment subtraction, without fitting any model. With the estimators

$$\hat{\lambda}_b \;:=\; \overline{\log V}_b - m_1, \qquad \hat{\sigma}_b^2 \;:=\; \max\big\{0,\; s_b^2(\log V) - m_2\big\},$$

where s_b^2 is the unbiased within-stratum sample variance, and with $R := 2(M + M_\lambda)$, each of

$$\big|\hat{\lambda}_b - \bar{\lambda}_b\big| \;\leq\; \delta_\lambda(n_b, \beta') \;:=\; 2(M+M_\lambda)\sqrt{\tfrac{\ln(2/\beta')}{2n_b}}, \qquad \big|\hat{\sigma}_b^2 - \sigma_b^2\big| \;\leq\; \delta_{\sigma^2}(n_b, \beta') \;:=\; R^2\sqrt{\tfrac{\ln(2/\beta')}{2n_b}}$$

fails with probability at most β'.

(b) Certificate. *Writing $\mathrm{Cov}_b(s_P) := \mathbb{Q}\big(V \leq s_P \,\big|\, b\big)$ and $\delta_\sigma := \sqrt{\hat{\sigma}_b^2 + \delta_{\sigma^2}} - \hat{\sigma}_b$, the stratum-$b$ zero-shot coverage obeys*

$$\mathrm{Cov}_b(s_P) \;\geq\; F_Z\Big(u\,e^{-\bar{\lambda}_b}\Big) \;-\; C_f\,u\,e^{-\bar{\lambda}_b}\cdot\sqrt{2}\,\sigma_b\,e^{\sigma_b^2}, \tag{18}$$

and the fully computable certificate

$$\mathrm{Cert}_b \;:=\; F_Z\Big(u\,e^{-(\hat{\lambda}_b+\delta_\lambda)}\Big) \;-\; C_f\,u\,e^{-(\hat{\lambda}_b-\delta_\lambda)}\cdot\sqrt{2}\,\big(\hat{\sigma}_b+\delta_\sigma\big)\,e^{(\hat{\sigma}_b+\delta_\sigma)^2} \tag{19}$$

satisfies, with probability at least $1-2K\beta'$ simultaneously over $b = 1, \ldots, K$,

$$\mathrm{Cov}_b(s_P) \;\geq\; \mathrm{Cert}_b. \tag{20}$$

Throughout this section, strata $\{S_b\}_{b=1}^K$ are fixed and X-measurable, and the n_b audit points falling in stratum b are i.i.d. draws from the conditional law $\mathbb{Q}(\cdot \mid X \in S_b)$ (this conditional-i.i.d. structure follows from i.i.d. sampling and X-measurability of the strata). We require $n_b \geq 1$ for the mean estimator and $n_b \geq 2$ for the variance estimator; strata not meeting these conditions are outside the scope of the simultaneous statement, as noted in the main text.

Proof of Theorem 4(a): moment identification and concentration. Step 1: moment identities. The decomposition equation 11 reads $\log V = \lambda(X) + W$ with $W = \log(Z/c)$. Under Assumption 1, $Z \perp X$ under $\mathbb{Q}$, and W is a measurable function of Z while $\lambda(X)$ is a measurable function of X; hence $\lambda(X) \perp W$, and the same independence holds conditionally on $X \in S_b$ (the conditioning event is X-measurable). Additivity of expectation gives $\mathbb{E}[\log V \mid b] = \bar{\lambda}_b + m_1$; additivity of variance under independence gives $\mathrm{Var}[\log V \mid b] = \mathrm{Var}[\lambda(X) \mid b] + \mathrm{Var}(W) = \sigma_b^2 + m_2$, where $m_1 = \mathbb{E}[W]$, $m_2 = \mathrm{Var}(W)$ depend only on the simulator shape (Assumption 2) and the pinball constant c, and are therefore simulator-computable.

Step 2: concentration of $\hat{\lambda}_b$. Under Assumption 4, $|\log V| = |\lambda(X) + W| \leq M + M_\lambda$ almost surely, so within stratum b the variables $\log V_i$ are i.i.d. and bounded in an interval of length $R = 2(M + M_\lambda)$; moreover $\mathbb{E}[\log V \mid$

$b] - m_1 = \bar{\lambda}_b$ by Step 1. Hoeffding's inequality (Lemma 2; conditions: independence, almost-sure boundedness) applied to $\hat{\lambda}_b = \overline{\log V}_b - m_1$ gives

$$\mathbb{Q}\Big(\big|\hat{\lambda}_b - \bar{\lambda}_b\big| \geq t\Big) \;\leq\; 2\exp\Big(-\frac{2n_b t^2}{R^2}\Big),$$

and the right side equals β' at

$$t \;=\; \delta_\lambda(n_b, \beta') \;=\; R\sqrt{\tfrac{\ln(2/\beta')}{2n_b}} \;=\; 2(M + M_\lambda)\sqrt{\tfrac{\ln(2/\beta')}{2n_b}}.$$

Step 3: concentration of $\hat{\sigma}_b^2$. Let s_b^2 denote the unbiased within-stratum sample variance of $\log V$, in the U-statistic form of Lemma 3(b); its expectation is $\mathrm{Var}[\log V \mid b] = \sigma_b^2 + m_2$. The bounded-difference constants are $c_i = R^2/n_b$ (Lemma 3(b); conditions: independence, range R), so McDiarmid's inequality gives

$$\mathbb{Q}\Big(\big|s_b^2 - (\sigma_b^2 + m_2)\big| \geq t\Big) \;\leq\; 2\exp\Big(-\frac{2n_b t^2}{R^4}\Big), \qquad t \;=\; \delta_{\sigma^2}(n_b, \beta') \;=\; R^2\sqrt{\tfrac{\ln(2/\beta')}{2n_b}}$$

at level β'. The truncation in $\hat{\sigma}_b^2 = \max\{0, s_b^2 - m_2\}$ does not enlarge the error: since $\sigma_b^2 \geq 0$ and $y \mapsto \max\{0, y\}$ is 1-Lipschitz with $\max\{0, \sigma_b^2\} = \sigma_b^2$,

$$\big|\hat{\sigma}_b^2 - \sigma_b^2\big| \;=\; \big|\max\{0, s_b^2 - m_2\} - \max\{0, \sigma_b^2\}\big| \;\leq\; \big|(s_b^2 - m_2) - \sigma_b^2\big| \;\leq\; \delta_{\sigma^2}$$

on the same event.

Step 4: from variance to standard deviation. Define $\hat{\sigma}_b := \sqrt{\hat{\sigma}_b^2}$ and $\delta_\sigma := \sqrt{\hat{\sigma}_b^2 + \delta_{\sigma^2}} - \hat{\sigma}_b \geq 0$. On the Step-3 event, $\sigma_b^2 \leq \hat{\sigma}_b^2 + \delta_{\sigma^2}$, hence

$$\sigma_b \;\leq\; \sqrt{\hat{\sigma}_b^2 + \delta_{\sigma^2}} \;=\; \hat{\sigma}_b + \delta_\sigma.$$

The rate distinction used below and in the main text is visible here: rationalizing, $\delta_\sigma = \delta_{\sigma^2}/\big(\sqrt{\hat{\sigma}_b^2 + \delta_{\sigma^2}} + \hat{\sigma}_b\big)$, so $\delta_\sigma \leq \delta_{\sigma^2}/(2\hat{\sigma}_b) = O(n_b^{-1/2})$ when $\hat{\sigma}_b$ is bounded away from zero, while near $\hat{\sigma}_b = 0$ one only has $\delta_\sigma = \sqrt{\delta_{\sigma^2}} = O(n_b^{-1/4})$, the square-root singularity of the variance-to-standard-deviation map at the origin. □

Proof of Theorem 4(b): the certificate. Step 1: population bound equation 18. By the key identity equation 4 applied conditionally on $X \in S_b$ with $s = s_P$ ($c\, s_P = u$),

$$\mathrm{Cov}_b(s_P) \;=\; \mathbb{E}\Big[F_Z\big(u\, e^{-\lambda(X)}\big)\,\Big|\, b\Big].$$

Decompose $\lambda(X) = \bar{\lambda}_b + \epsilon_b(X)$ with $\epsilon_b(X) := \lambda(X) - \bar{\lambda}_b$ centered and σ_b-sub-Gaussian under the conditional law (Assumption 3). Inserting $F_Z\big(ue^{-\bar{\lambda}_b}\big)$ and using the Lipschitz bound $|F_Z(a) - F_Z(b)| \leq C_f|a - b|$ pointwise,

$$\big|\mathrm{Cov}_b(s_P) - F_Z\big(ue^{-\bar{\lambda}_b}\big)\big| \;\leq\; C_f\, u\, \mathbb{E}\Big[\,\big|e^{-\lambda(X)} - e^{-\bar{\lambda}_b}\big|\,\Big|\, b\Big] \;=\; C_f\, u\, e^{-\bar{\lambda}_b}\, \mathbb{E}\Big[\,\big|e^{-\epsilon_b(X)} - 1\big|\,\Big|\, b\Big],$$

where the second step factors out the deterministic $e^{-\bar{\lambda}_b}$. Lemma 5(iii) (applicable by Assumption 3: centeredness and the mgf bound) controls the last expectation by $\sqrt{2}\,\sigma_b\, e^{\sigma_b^2}$, giving equation 18:

$$\mathrm{Cov}_b(s_P) \;\geq\; F_Z\big(u\, e^{-\bar{\lambda}_b}\big) \;-\; C_f\, u\, e^{-\bar{\lambda}_b}\, \sqrt{2}\,\sigma_b\, e^{\sigma_b^2}.$$

Step 2: term-by-term comparison on the concentration event. For stratum b, let $\mathcal{E}_b$ be the intersection of the two events of Theorem 4(a), $\mathbb{Q}(\mathcal{E}_b) \geq 1 - 2\beta'$:

$$\big|\hat{\lambda}_b - \bar{\lambda}_b\big| \;\leq\; \delta_\lambda, \qquad \sigma_b \;\leq\; \hat{\sigma}_b + \delta_\sigma.$$

On $\mathcal{E}_b$ we compare equation 18 with equation 19 term by term. *First term.* The map $\ell \mapsto F_Z\big(u\, e^{-\ell}\big)$ is nonincreasing (F_Z nondecreasing, $e^{-\ell}$ decreasing, $u > 0$), and $\bar{\lambda}_b \leq \hat{\lambda}_b + \delta_\lambda$ on $\mathcal{E}_b$; hence

$$F_Z\big(u\, e^{-\bar{\lambda}_b}\big) \;\geq\; F_Z\Big(u\, e^{-(\hat{\lambda}_b + \delta_\lambda)}\Big).$$

Second (subtracted) term. This term must be *upper*-bounded. On $\mathcal{E}_b$, $\bar{\lambda}_b \geq \hat{\lambda}_b - \delta_\lambda$, so $e^{-\bar{\lambda}_b} \leq e^{-(\hat{\lambda}_b - \delta_\lambda)}$. The function $h(\sigma) := \sqrt{2}\,\sigma\, e^{\sigma^2}$ is nondecreasing on $\sigma \geq 0$, since $h'(\sigma) = \sqrt{2}\, e^{\sigma^2}\big(1 + 2\sigma^2\big) > 0$; combined with $\sigma_b \leq \hat{\sigma}_b + \delta_\sigma$ this gives $h(\sigma_b) \leq h\big(\hat{\sigma}_b + \delta_\sigma\big)$. Multiplying the nonnegative factors,

$$C_f\, u\, e^{-\bar{\lambda}_b}\, \sqrt{2}\,\sigma_b\, e^{\sigma_b^2} \;\leq\; C_f\, u\, e^{-(\hat{\lambda}_b - \delta_\lambda)}\, \sqrt{2}\,\big(\hat{\sigma}_b + \delta_\sigma\big)\, e^{(\hat{\sigma}_b + \delta_\sigma)^2}.$$

Subtracting a larger correction and lower-bounding the leading term yields $\mathrm{Cov}_b(s_P) \geq \mathrm{Cert}_b$ on $\mathcal{E}_b$.

Step 3: simultaneity. $\mathbb{Q}\big(\cap_{b=1}^{K} \mathcal{E}_b\big) \geq 1 - 2K\beta'$ by the union bound over K strata of the two per-stratum events of level β' each; on the intersection, Step 2 holds for every b, proving equation 20. □

Proof of the zero-tightness and consistency properties. Zero-tightness. Suppose the simulator is perfect, $\lambda \equiv 0$. Then $\bar{\lambda}_b = 0$ and $\sigma_b = 0$, and equation 18 reads $\mathrm{Cov}_b(s_P) \geq F_Z(u) = 1 - \alpha$; since in fact $\mathrm{Cov}_b(s_P) = F_Z(u)$ exactly (identity equation 4 with $\lambda \equiv 0$), the population bound is tight at the zero-degradation point. For the computable certificate, Steps 2–4 of part (a) give $\hat{\lambda}_b \to 0$, $\hat{\sigma}_b \to 0$ in probability as $n_b \to \infty$, and $\delta_\lambda, \delta_{\sigma^2}, \delta_\sigma \to 0$; both terms of equation 19 are continuous in these quantities, so $\mathrm{Cert}_b \to F_Z(u) = 1 - \alpha$: the certificate is asymptotically tight at zero degradation and does not alarm indefinitely.

Consistency. As $n_b \to \infty$ with β' fixed, $\delta_\lambda, \delta_{\sigma^2} = O(n_b^{-1/2})$ and $(\hat{\lambda}_b, \hat{\sigma}_b) \to (\bar{\lambda}_b, \sigma_b)$; hence Cert_b converges to the right side of equation 18, a legitimate lower bound on the true stratum coverage. The convergence rate of the estimation slack is governed by δ_σ: when σ_b is bounded away from zero, $\delta_\sigma \leq \delta_{\sigma^2}/(2\hat{\sigma}_b) = O(n_b^{-1/2})$ and all estimation errors in equation 19 vanish at the $O(n_b^{-1/2})$ rate (the map $(\hat{\lambda}, \hat{\sigma}) \mapsto \mathrm{Cert}_b$ is locally Lipschitz there); near $\sigma_b = 0$, $\delta_\sigma = \sqrt{\delta_{\sigma^2}} = O(n_b^{-1/4})$ dominates, and the slack vanishes only at the $O(n_b^{-1/4})$ rate, because the standard-deviation estimate has a square-root singularity at zero variance while the correction term $h(\sigma) = \sqrt{2}\sigma e^{\sigma^2}$ has nonvanishing slope at $\sigma = 0$. □

Remark 6 (Identifiability boundary of moment subtraction). Moment subtraction identifies the first two moments of the law of λ, not the law itself: arbitrary tail-shape changes are not dominated by a second-moment correction, which is why the audit of Proposition 4 is not optional.

Remark 7 (Computability budget). Every ingredient of equation 19 is computable: F_Z, u, m_1, m_2 from simulator resampling at Monte Carlo precision (resampling error can be folded into the δ-budget at the price of additional union-bound terms), $\hat{\lambda}_b, \hat{\sigma}_b$ from $n_b \geq 2$ audit points, and the δ-terms in closed form. The single exception is $C_f = \sup_z f_Z(z)$, which requires an analytic upper bound or density regularity: no resampling scheme recovers a sup-norm without assumptions. Whether Cert_b is nontrivial at small n_b depends on the concrete constants $M, M_\lambda, C_f, u, \beta', K$; this is an empirical property, not a theorem.

K A Convexity-Verified Jensen Variant of the Certificate

The certificate equation 19 pays the correction $\sqrt{2}\,\sigma_b\, e^{\sigma_b^2}$ for within-stratum dispersion of λ, inherited from the Lipschitz-plus-mgf route of Theorem 4(b). When the map $g(t) := F_Z(u\,e^{-t})$ is convex on the relevant region, Jensen's inequality removes this correction entirely. The variant below is what the deployed certificate of Section 5.4 implements: convexity is a purely simulator-side, numerically verifiable property (Remark 8), the certificate fires only when the verification passes, and falls back to equation 19 otherwise.

Proposition 3 (Convexity-verified Jensen certificate)**.** *Assume the setting of Theorem 4 (Assumptions 1, 2, 3, 4), write $u = c\,s_P$, and define*

$$g(t) \;:=\; F_Z\big(u\,e^{-t}\big), \qquad t \in \mathbb{R}. \tag{21}$$

Let $\mathcal{E}_b$ and $\delta_\lambda, \delta_\sigma$ be as in Theorem 4.

(i) Convexity on the moment support. *If g is convex on $[-M_\lambda, M_\lambda]$, then on $\cap_b \mathcal{E}_b$ — an event of probability at least $1 - 2K\beta'$, the same event as in Theorem 4, so no additional probability budget is spent — simultaneously for $b = 1, \ldots, K$,*

$$\mathrm{Cov}_b(s_P) \;\geq\; \mathrm{Cert}_b^{\mathrm{J}} \;:=\; F_Z\Big(u\,e^{-(\hat{\lambda}_b + \delta_\lambda)}\Big), \tag{22}$$

i.e. the entire σ-correction of equation 19 vanishes.

(ii) Verified-interval version. *Suppose convexity is verified only on the computable interval*

$$I_b \;:=\; \big[\hat{\lambda}_b - \delta_\lambda - \tau_b,\;\; \hat{\lambda}_b + \delta_\lambda + \tau_b\big], \qquad \tau_b \;:=\; \kappa\,\big(\hat{\sigma}_b + \delta_\sigma\big),\;\; \kappa \geq 1,$$

and define the tail-mass bound and the local Lipschitz constant

$$\bar{p}_b \;:=\; 2\exp\Big(-\frac{\tau_b^2}{2(\hat{\sigma}_b + \delta_\sigma)^2}\Big), \qquad L_b \;:=\; C_f\,u\,e^{-(\hat{\lambda}_b - \delta_\lambda - \tau_b)}$$

(with the convention $\bar{p}_b = 0$ when $\hat{\sigma}_b + \delta_\sigma = 0$: then $\sigma_b = 0$ on $\mathcal{E}_b$, $\lambda = \bar{\lambda}_b$ a.s. in stratum b, and the statement reduces to part (i) on I_b). Then on $\cap_b \mathcal{E}_b$, simultaneously for all b,

$$\mathrm{Cov}_b(s_P) \;\geq\; \mathrm{Cert}_b^{\mathrm{J,tr}} \;:=\; \big(1 - \bar{p}_b\big)\, F_Z\Big(u\, e^{-(\hat{\lambda}_b + \delta_\lambda)}\Big) \;-\; 2M_\lambda\, L_b\, \bar{p}_b. \tag{23}$$

Whenever the convexity verification fails, the certificate falls back to equation 19; since both bounds lower-bound $\mathrm{Cov}_b(s_P)$ on the same event $\cap_b \mathcal{E}_b$, the pointwise maximum of the firing rule and the fallback is again a valid certificate at the same confidence $1 - 2K\beta'$.

Remark 8 (The convexity criterion is simulator-checkable). With $z := u\, e^{-t}$, differentiating equation 21 twice gives

$$g'(t) \;=\; -z\, f_Z(z), \qquad g''(t) \;=\; z\,\big(f_Z(z) + z\, f_Z'(z)\big) \;=\; z\,\big(z\, f_Z(z)\big)',$$

so, since $z > 0$ and $t \mapsto ue^{-t}$ is a decreasing bijection,

$$g \text{ convex on } I \iff z \mapsto z\, f_Z(z) \text{ nondecreasing on } \{ue^{-t} : t \in I\} \iff \big(z\, f_Z(z)\big)' \;\geq\; 0 \text{ there.}$$

The rightmost condition involves only the simulator's own density f_Z and is checked numerically on a grid over the image of I_b; no sup-norm or tail assumption on f_Z is needed. Grid-verification error is a simulator-side numerical issue of the same nature as the Monte Carlo error in F_Z itself and can be folded into the δ-budget exactly as in the computability remark of Theorem 4.

Proof of Proposition 3. (i) By the key identity equation 4 applied conditionally on $X \in S_b$,

$$\mathrm{Cov}_b(s_P) \;=\; \mathbb{E}\big[g\big(\lambda(X)\big) \,\big|\, b\big].$$

Under Assumption 4, $|\lambda(X)| \leq M_\lambda$ almost surely, hence also conditionally on b, so the conditional law of $\lambda(X)$ given b is supported on the interval $[-M_\lambda, M_\lambda]$ where g is convex; Jensen's inequality therefore applies and gives

$$\mathrm{Cov}_b(s_P) \;\geq\; g\big(\bar{\lambda}_b\big).$$

On $\mathcal{E}_b$, $\bar{\lambda}_b \leq \hat{\lambda}_b + \delta_\lambda$, and g is nonincreasing (the composition of the nondecreasing F_Z with the decreasing map $t \mapsto ue^{-t}$), hence $g(\bar{\lambda}_b) \geq g\big(\hat{\lambda}_b + \delta_\lambda\big) = \mathrm{Cert}_b^{\mathrm{J}}$.

(ii) Fix a stratum b and work on $\mathcal{E}_b$, writing $p := \mathbb{Q}\big(\lambda(X) \notin I_b \,\big|\, b\big)$. Since $|\hat{\lambda}_b - \bar{\lambda}_b| \leq \delta_\lambda$ on $\mathcal{E}_b$, the interval I_b contains $[\bar{\lambda}_b - \tau_b,\, \bar{\lambda}_b + \tau_b]$, and the sub-Gaussian mgf of Assumption 3 with the Chernoff method gives

$$p \;\leq\; \mathbb{Q}\Big(\big|\lambda(X) - \bar{\lambda}_b\big| > \tau_b \,\Big|\, b\Big) \;\leq\; 2\exp\Big(-\frac{\tau_b^2}{2\sigma_b^2}\Big) \;\leq\; \bar{p}_b,$$

where the last step uses $\sigma_b \leq \hat{\sigma}_b + \delta_\sigma$ on $\mathcal{E}_b$ and monotonicity of the bound in σ_b. Decomposing the key identity at I_b and using $g \geq 0$,

$$\mathrm{Cov}_b(s_P) \;\geq\; (1-p)\,\mathbb{E}\big[g\big(\lambda(X)\big) \,\big|\, b,\; \lambda(X) \in I_b\big] \;\geq\; (1-p)\, g(\mu_I), \qquad \mu_I \;:=\; \mathbb{E}\big[\lambda(X) \,\big|\, b,\; \lambda(X) \in I_b\big],$$

where the second step is Jensen applied to the conditional law given $\{b,\, \lambda(X) \in I_b\}$, which is supported on the interval I_b on which g is convex. Write $\nu := \mathbb{E}\big[\lambda(X) \,\big|\, b,\; \lambda(X) \notin I_b\big]$; the mean identity $\bar{\lambda}_b = (1-p)\,\mu_I + p\,\nu$ gives $\bar{\lambda}_b - \mu_I = p\,(\nu - \mu_I)$, and since $|\lambda(X)| \leq M_\lambda$ almost surely,

$$|\mu_I - \bar{\lambda}_b| \;\leq\; 2\, M_\lambda\, p \;\leq\; 2\, M_\lambda\, \bar{p}_b.$$

Both μ_I and $\bar{\lambda}_b$ lie in I_b (for $\bar{\lambda}_b$, by $|\hat{\lambda}_b - \bar{\lambda}_b| \leq \delta_\lambda$), and g is L_b-Lipschitz on I_b because $|g'(t)| = z\, f_Z(z) \leq C_f\, u\, e^{-t} \leq L_b$ for all $t \geq \hat{\lambda}_b - \delta_\lambda - \tau_b$. Since $\bar{\lambda}_b \leq \hat{\lambda}_b + \delta_\lambda$ and g is nonincreasing,

$$g(\mu_I) \;\geq\; g(\bar{\lambda}_b) \;-\; L_b\,|\mu_I - \bar{\lambda}_b| \;\geq\; F_Z\Big(u\, e^{-(\hat{\lambda}_b + \delta_\lambda)}\Big) \;-\; 2M_\lambda\, L_b\, \bar{p}_b.$$

Multiplying by $1 - p \geq 1 - \bar{p}_b$ (the leading term is nonnegative) and relaxing $(1-p)$ to 1 in the subtracted term yields equation 23. Both parts hold on $\cap_b \mathcal{E}_b$, whose probability is at least $1 - 2K\beta'$ by Theorem 4. □

Remark 9 (Properties of the Jensen certificate). *(a) Zero-tightness where verification passes.* When the check passes, the subtracted σ-correction is absent, so $\mathrm{Cert}_b^{\mathrm{J}}$ is strictly tighter than equation 19 whenever it fires; at the zero-degradation point $\lambda \equiv 0$, $\mathrm{Cert}_b^{\mathrm{J}} \to F_Z(u) = 1 - \alpha$ as $n_b \to \infty$, and the same limit holds for equation 23 since $\tau_b, \bar{p}_b \to 0$. *(b) Generic obstruction, and the noise-family escape.* The check is local around the mean, but the

effective argument $z = u\,e^{-t}$ for t near $\bar{\lambda}_b$ brackets $u = Q_{1-\alpha}(Z)$, and any density that decreases in its upper tail has $(z f_Z(z))' < 0$ somewhere past the quantile. The verification interval I_b — which must cover the confidence width δ_λ plus a $\kappa\hat{\sigma}_b$ margin around the unknown $\bar{\lambda}_b$ — then generically straddles the edge where convexity fails, and the certificate falls back; this is a property of tail geometry, not of the audit size. The obstruction disappears for a bounded family whose density increases throughout its support (Beta-type): there $z\,f_Z(z)$ is nondecreasing on the entire support, so the check passes at the nominal point at every audit size and part (a) applies. Section 5.4 uses such a family and observes the predicted zero-tightness. *(c) The Jensen gap.* Even when it fires, the certificate lower-bounds $g(\bar{\lambda}_b)$ rather than $\mathbb{E}[g(\lambda(X)) \mid b]$; the deficit $\mathbb{E}[g(\lambda(X)) \mid b] - g(\bar{\lambda}_b) \approx \frac{1}{2}\, g''(\bar{\lambda}_b)\, \sigma_b^2$ is the price of dropping the σ-correction, and remains a valid lower bound by Jensen. *(d) Firing rule and constants.* The deployed certificate uses $\kappa = 4$; then $\bar{p}_b \le 2e^{-8} \approx 6.7 \times 10^{-4}$, so both correction terms in equation 23 are negligible relative to the leading term, which is what makes the verified-interval certificate practically indistinguishable from the ideal equation 22 whenever the check passes.

L Statement and Proof of Proposition 4

We restate the proposition in full.

Proposition 4 (Shape-compatibility audit)**.** *In stratum b, let $\tilde{F}_b$ be the model-implied law of V obtained by Monte Carlo convolution of the moment-matched parametric family $\lambda \sim \mathcal{N}(\hat{\lambda}_b, \hat{\sigma}_b^2)$ (the natural maximum-entropy choice under Assumption 3) with simulator resamples of Z/c. Let $D_b := \sup_t \left|\hat{F}_{n_b}(t) - \tilde{F}_b(t)\right|$ be the Kolmogorov–Smirnov statistic against the empirical cdf, with critical value from a parametric bootstrap at level γ_b ($\sum_b \gamma_b = \gamma$). If* within each bootstrap replicate *the parameters $\hat{\lambda}_b, \hat{\sigma}_b^2$ are re-estimated and $\tilde{F}_b$ reconstructed before recomputing D_b, and the Monte Carlo error in $\tilde{F}_b$ is controlled, then under standard regularity conditions the test has asymptotic level γ_b, and its power tends to one against any fixed alternative outside the limiting fitted model.*

We work in a fixed stratum b with $n := n_b$ audit points, and write the observed variable as $Y = \log V$. The fitted parametric family is $\{F_\theta : \theta = (\mu, s^2) \in \Theta \subset \mathbb{R} \times (0, \infty)\}$, where F_θ is the law of $\lambda + \log(Z/c)$ (the log-space convolution) with $\lambda \sim \mathcal{N}(\mu, s^2)$ and Z/c resampled from the simulator; this matches the observation space, and the moment estimators $\hat{\theta} = (\hat{\lambda}_b, \hat{\sigma}_b^2)$ are the natural log-space estimators of Theorem 4(a). Set $\tilde{F}_b = F_{\hat{\theta}}$ and $D_b = \sup_t \left|\hat{F}_n(t) - F_{\hat{\theta}}(t)\right|$. (Equivalently, the audit could be run on the scale of V against the law of $e^\lambda(Z/c)$: the KS distance is invariant under the strictly increasing transform $y \mapsto e^y$, so the two formulations coincide and the log-space choice is what makes the moment estimators natural.) The argument uses the following regularity conditions, which we state explicitly:

- (R1) the audit points $Y_1, \ldots, Y_n$ are i.i.d. from a law F (true law, equal to F_{θ_0} for some interior $\theta_0 \in \Theta$ under the composite null);
- (R2) $\theta \mapsto F_\theta$ is differentiable at θ_0 in sup norm: $\sup_t \left|F_\theta(t) - F_{\theta_0}(t) - \nabla_\theta F_{\theta_0}(t)^\top(\theta - \theta_0)\right| = o\big(|\theta - \theta_0|\big)$, with $t \mapsto \nabla_\theta F_{\theta_0}(t)$ bounded;
- (R3) $\hat{\theta}$ is consistent for its population target and asymptotically linear under the null: $\sqrt{n}\,(\hat{\theta} - \theta_0) = n^{-1/2}\sum_i \psi(Y_i) + o_p(1)$ with $\mathbb{E}\psi = 0$, $\mathbb{E}\|\psi\|^2 < \infty$ (for the moment estimator $(\hat{\lambda}_b, \hat{\sigma}_b^2)$ this holds under finite fourth moments of Y, which Assumption 4 provides);
- (R4) F_{θ_0} is continuous (implied by Assumption 2, since F_θ is a mixture of scaled copies of the continuous F_Z);
- (R5) the Monte Carlo error in constructing $\tilde{F}_b$ and the bootstrap replicates is controlled: the convolution F_θ is evaluated with M_{mc} simulator draws, whose sup-norm error is $O_p(M_{\mathrm{mc}}^{-1/2})$ by the DKW inequality applied to the Monte Carlo empirical cdf, and M_{mc} is taken $\gg n$ so that this error is $o_p(n^{-1/2})$.

Proof of Proposition 4. Step 1: null limit of the statistic. Under the composite null (R1 with $F = F_{\theta_0}$), decompose

$$\sqrt{n}\,\big(\hat{F}_n(t) - F_{\hat{\theta}}(t)\big) \;=\; \sqrt{n}\,\big(\hat{F}_n(t) - F_{\theta_0}(t)\big) \;-\; \sqrt{n}\,\big(F_{\hat{\theta}}(t) - F_{\theta_0}(t)\big).$$

The class of indicators $\{\mathbb{I}(Y \le t) : t \in \mathbb{R}\}$ is Donsker, so the first term converges weakly (as a process in t) to a Brownian bridge $\mathbb{B}\big(F_{\theta_0}(t)\big)$; by (R2)–(R3) the second term equals $\nabla_\theta F_{\theta_0}(t)^\top\, \sqrt{n}(\hat{\theta} - \theta_0) + o_p(1)$ uniformly in t and converges jointly with the first. Hence $\sqrt{n}\,\big(\hat{F}_n - F_{\hat{\theta}}\big)$ converges weakly to a centered Gaussian process

$$\mathbb{G}(t) \;=\; \mathbb{B}\big(F_{\theta_0}(t)\big) \;-\; \nabla_\theta F_{\theta_0}(t)^\top \mathbb{Z}, \qquad \mathbb{Z} \sim \mathcal{N}\big(0, \mathbb{E}[\psi\psi^\top]\big),$$

with $\mathbb{B}$ and $\mathbb{Z}$ jointly Gaussian (the Durbin/Darling estimated-parameter limit; continuity of F_{θ_0} in (R4) ensures the bridge is well-defined and the sup functional is continuous). By the continuous mapping theorem, $\sqrt{n}\, D_b \Rightarrow \sup_t |\mathbb{G}(t)|$.

Step 2: validity of the re-estimating parametric bootstrap. In the bootstrap world, conditional on the data, replicates $Y_1^*, \ldots, Y_n^*$ are drawn i.i.d. from $F_{\hat\theta}$; the parameters $\hat\theta^* = (\hat\lambda_b^*, \hat\sigma_b^{2*})$ are *re-estimated* from the replicate and $F_{\hat\theta^*}$ is reconstructed before computing $D_b^* = \sup_t |\hat F_n^*(t) - F_{\hat\theta^*}(t)|$. The bootstrap world satisfies the null by construction, with $\hat\theta$ playing the role of θ_0; since $\hat\theta \to \theta_0$ and $\sup_t |F_{\hat\theta}(t) - F_{\theta_0}(t)| \to 0$ in probability (R2)–(R3), the same decomposition as Step 1 applied along the triangular array of bootstrap laws, together with the uniform (in a shrinking neighborhood of θ_0) versions of the Donsker and differentiability conditions, which (R2)–(R3) supply, gives conditional weak convergence $\sqrt{n}\,(\hat F_n^* - F_{\hat\theta^*}) \Rightarrow \mathbb{G}$ in probability, and hence $\sqrt{n}\, D_b^* \Rightarrow \sup_t |\mathbb{G}(t)|$ in probability. The bootstrap critical value $c_n(\gamma_b)$, the $(1-\gamma_b)$-quantile of the replicate distribution, therefore converges in probability to the $(1-\gamma_b)$-quantile of $\sup_t |\mathbb{G}(t)|$, which is a continuity point of that law under the stated Gaussian non-degeneracy; it follows that

$$\mathbb{Q}\big(D_b > c_n(\gamma_b)\big) \;\longrightarrow\; \gamma_b \qquad \text{under the composite null,}$$

i.e. the test has asymptotic level γ_b. The allocation $\sum_b \gamma_b = \gamma$ then controls the familywise level across strata by the union bound.

Step 3: why re-estimation is necessary. The naive bootstrap, which treats the fitted parameters as known and compares $\hat F_n^*$ with the fixed $F_{\hat\theta}$, reproduces instead the limit $\mathbb{B}(F_{\theta_0}(t))$: the estimation-drift term $\nabla_\theta F_{\theta_0}(t)^\top \mathbb{Z}$ is absent. The laws of $\sup_t |\mathbb{B}(F_{\theta_0}(t))|$ and $\sup_t |\mathbb{G}(t)|$ differ in general (the latter has smaller pointwise variance by the subtraction), so the naive critical value converges to the wrong quantile and the naive test does not have asymptotic level γ_b. Re-estimation inside each replicate is precisely what restores the estimated-parameter null law; the Monte Carlo layer is absorbed by (R5).

Step 4: power (KS consistency). Let F be a fixed alternative with $\kappa := \inf_{\theta\in\Theta} \sup_t |F(t) - F_\theta(t)| > 0$ (any law outside the closure of the fitted family, which includes the composite null: $\lambda \mid b$ Gaussian, $\lambda \perp W$, W of simulator shape). The moment estimator converges to its population moment target $\theta^\dagger \in \Theta$ (consistency of method-of-moments limits under (R3)'s moment conditions), so $\sup_t |F_{\hat\theta}(t) - F_{\theta^\dagger}(t)| = o_p(1)$ by (R2). The class of half-lines is Glivenko–Cantelli, hence $\sup_t |\hat F_n(t) - F(t)| = o_p(1)$, and therefore

$$D_b \;\geq\; \sup_t \big|F(t) - F_{\hat\theta}(t)\big| - \sup_t \big|\hat F_n(t) - F(t)\big| \;\geq\; \kappa - o_p(1).$$

Meanwhile the bootstrap world is null by construction, so by Step 2 its statistic is $O_p(n^{-1/2})$ and the critical value satisfies $c_n(\gamma_b) = O_p(n^{-1/2})$. Consequently $\mathbb{Q}\big(D_b > c_n(\gamma_b)\big) \to 1$: the power tends to one against any fixed alternative outside the limiting fitted model. □

Remark 10 (Semantics of the composite null). The null audited above is composite: $\lambda \mid b$ Gaussian *and* $\lambda(X) \perp W$ *and* W distributed as the simulator shape. Rejection cannot distinguish which component failed and must be read as a generic model-mismatch alarm, triggering the Theorem 3(a) regime and full Stage-II recalibration of the stratum; non-rejection does not validate the certificate equation 19, since a goodness-of-fit test cannot prove its null; it reports only that no mismatch was detected at the current sample size and power.

M Statement and Proof of Theorem 5

We restate the theorem in full.

Theorem 5 (Stratified exact coverage, on average)**.** *Let the calibration points $\{(X_i, V_i)\}$ be i.i.d. and the test point exchangeable with them. Then for every stratum b,*

$$\mathbb{Q}\big(V \leq \hat s_{b(X)} \,\big|\, X \in S_b\big) \;\geq\; 1 - \alpha, \tag{24}$$

with probability taken over the joint randomness of the calibration set and the test point.

Proof of Theorem 5. We use the standing conventions of the main text: the partition $\{S_b\}_{b=1}^K$ is determined by a fixed X-measurable function with $\mathbb{Q}(X \in S_b) > 0$ for every retained stratum; the calibration points $\{(X_i, V_i)\}_{i=1}^{n_{\mathrm{cal}}}$ are i.i.d. from $\mathbb{Q}$; the test point (X, V) is exchangeable with them; $k_b = \lceil (n_b+1)(1-\alpha) \rceil$ with the convention $V_{(n_b+1)} := +\infty$ (so coverage is trivially 1 when $k_b = n_b + 1$, including $n_b = 0$); and ties are broken at random.

Step 1: exchangeability within a stratum. Fix a stratum b and condition on the event $\{X \in S_b\}$ and on the count $n_b = m \geq 0$ of calibration points falling in S_b. Because stratum membership is a measurable function of each point's own covariate, the selection indicators $\mathbb{I}\{X_i \in S_b\}$ are functions of the i.i.d. pairs (X_i, V_i); conditioning on the X-measurable configuration $\{X \in S_b,\ n_b = m\}$ (an event measurable with respect to the covariates alone) therefore preserves exchangeability of the selected pairs: the m within-stratum calibration scores together with the test score V are $m+1$ exchangeable random variables, each distributed as $\mathbb{Q}(\cdot \mid X \in S_b)$. Formally, for any bounded test function ϕ and any permutation π of the $m+1$ scores, $\mathbb{E}\big[\phi \mid X \in S_b, n_b = m\big]$ is invariant under π because the joint law of $\big((X_i, V_i), (X, V)\big)$ is i.i.d. and the conditioning event is invariant under permutations of the selected indices.

Step 2: rank uniformity. Among $m+1$ exchangeable scores with ties broken at random, the rank of V is uniformly distributed on $\{1, \ldots, m+1\}$ (symmetry: each ordering of the exchangeable vector is equally likely, and random tie-breaking makes the rank exactly uniform even with atoms). Consequently, for $k_b \leq m$,

$$\mathbb{Q}\big(V \leq V_{(k_b)} \,\big|\, X \in S_b,\ n_b = m\big) \ \geq\ \frac{k_b}{m+1} \ =\ \frac{\lceil (m+1)(1-\alpha) \rceil}{m+1} \ \geq\ 1-\alpha,$$

and for $k_b = m+1$ the convention $V_{(m+1)} = +\infty$ makes the left side $1 \geq 1-\alpha$. (Without randomized tie-breaking the standard conformal inequality is conservative in the same direction, so the statement is unaffected.)

Step 3: averaging over the random count. The bound of Step 2 holds for every value m of n_b, hence

$$\mathbb{Q}\big(V \leq \hat{s}_{b(X)} \,\big|\, X \in S_b\big) \ =\ \mathbb{E}\Big[\mathbb{Q}\big(V \leq V_{(k_b)} \,\big|\, X \in S_b,\ n_b\big) \,\Big|\, X \in S_b\Big] \ \geq\ 1-\alpha,$$

where the outer expectation averages over the random n_b (whose conditional law given $X \in S_b$ is unconstrained; in particular, no minimum sample size is required). This is equation 24, with probability over the joint randomness of the calibration set and the test point. No model assumption (not even Assumption 1) entered: the argument uses only exchangeability and the X-measurability of the strata. If the strata were estimated from the same calibration data, the selection event would no longer be covariate-measurable independent of the scores, Step 1 would fail, and sample splitting or a re-proved exchangeable selection rule would be required. □

N Statement and Proof of Theorem 6

We restate the theorem in full.

Theorem 6 (Realized uniform guarantee and sample complexity)**.** *Let $G_b(t) := \mathbb{Q}(V \leq t \mid X \in S_b)$ be continuous. Then with probability at least $1-\beta$ over the calibration set,*

$$G_b(\hat{s}_b) \ \geq\ 1 - \alpha - \epsilon_{n_b}, \qquad \epsilon_{n_b} \ :=\ \sqrt{\tfrac{\ln(2K/\beta)}{2(n_b+1)}}, \quad \textit{simultaneously for all } b = 1, \ldots, K. \tag{25}$$

Consequently, realized slack at most δ in every stratum is guaranteed by the sufficient condition $n_b \geq \ln(2K/\beta)/\big(2(\delta - 1/(n_b+1))^2\big)$ per stratum, with dominant term $\ln(2K/\beta)/(2\delta^2)$, and, under balanced allocation $n_b = n_{\mathrm{cal}}/K$, by

$$n_{\mathrm{cal}} \ \geq\ \frac{K\,\ln(2K/\beta)}{2\delta^2}\,\big(1 + o(1)\big) \ =\ O\big(K \log K \,/\, \delta^2\big). \tag{26}$$

Proof of Theorem 6. Step 1: per-stratum tail bound. Fix a stratum b and write $n := n_b$. Conditional on the stratum counts, the n within-stratum calibration scores are i.i.d. from $\mathbb{Q}(\cdot \mid X \in S_b)$ with cdf G_b, which is continuous by hypothesis; these are exactly the applicability conditions of Lemma 4 (independence, continuous cdf). If $k_b = \lceil (n+1)(1-\alpha) \rceil = n+1$ then $\hat{s}_b = +\infty$ by convention, $G_b(\hat{s}_b) = 1$, and the statement holds trivially; assume $k_b \leq n$. Then $U := G_b(\hat{s}_b) \sim \mathrm{Beta}(k_b, n+1-k_b)$ and the Beta–binomial identity (Lemma 4(b)) gives, for $q \in (0,1)$,

$$\mathbb{Q}\big(U \leq q\big) \ =\ \mathbb{Q}\big(\mathrm{Bin}(n,q) \geq k_b\big).$$

Set $c_b := k_b/(n+1) \geq 1-\alpha$ (the ceiling) and $q = c_b - t$ with $0 < t \leq c_b$. The binomial mean is $nq = nc_b - nt$, so the mean-to-threshold gap is

$$k_b - nq \ =\ (n+1)\,c_b - nc_b + nt \ =\ c_b + nt \ \geq\ (n+1)\,t,$$

using $c_b \geq t$. Hoeffding's inequality (Lemma 2 applied to the sum of n independent Bernoulli(q) variables, range 1; conditions: independence, boundedness) yields

$$\mathbb{Q}\big(U \leq c_b - t\big) \ \leq\ \exp\Big(-\frac{2\,(c_b + nt)^2}{n}\Big) \ \leq\ \exp\big(-2\,(n+1)^2 t^2/n\big) \ \leq\ \exp\big(-2\,(n+1)\,t^2\big),$$

since $(n+1)^2/n \geq n+1$. (For $t > c_b$ the event is empty and the display is trivially valid.) Taking

$$t \;=\; \epsilon_{n_b} \;:=\; \sqrt{\frac{\ln(2K/\beta)}{2\,(n_b+1)}}$$

makes each stratum's failure probability at most $\exp\big(-\ln(2K/\beta)\big) = \beta/(2K)$. On the complementary per-stratum event,

$$G_b(\hat{s}_b) \;>\; c_b - \epsilon_{n_b} \;\geq\; 1-\alpha-\epsilon_{n_b},$$

where the ceiling inequality $c_b \geq 1-\alpha$ absorbs the discretization of k_b exactly: *no additional* $1/(n_b+1)$ *slack term appears*.

Step 2: union bound. The events of Step 1 depend only on the disjoint within-stratum subsamples; applying the union bound over the K strata (valid regardless of dependence),

$$\mathbb{Q}\Big(G_b(\hat{s}_b) \;\geq\; 1-\alpha-\epsilon_{n_b} \text{ for all } b=1,\dots,K\Big) \;\geq\; 1-K\cdot\frac{\beta}{2K} \;=\; 1-\frac{\beta}{2} \;\geq\; 1-\beta,$$

which is equation 25.

Step 3: sample complexity equation 26. To guarantee realized slack at most δ in every stratum it suffices that $\epsilon_{n_b} \leq \delta$, i.e. $n_b+1 \geq \ln(2K/\beta)/(2\delta^2)$. A convenient implicit sufficient condition, slightly stronger, which also bounds the *two-sided* deviation $|G_b(\hat{s}_b)-(1-\alpha)|$ (since the ceiling contributes at most $c_b-(1-\alpha) < 1/(n_b+1)$ on the upper side), is obtained by requiring $\epsilon_{n_b} \leq \delta - 1/(n_b+1)$, equivalently

$$n_b \;\geq\; \frac{\ln(2K/\beta)}{2\,\big(\delta-1/(n_b+1)\big)^2} \;-\; 1,$$

whose dominant term is $\ln(2K/\beta)/(2\delta^2)$ as soon as $n_b+1 \geq 2/\delta$; the main text states the slightly stronger (still sufficient) form $n_b \geq \ln(2K/\beta)/\big(2(\delta-1/(n_b+1))^2\big)$ without the trailing -1, which only enlarges the required n_b by one. Under balanced allocation $n_b = n_{\mathrm{cal}}/K$ this becomes

$$n_{\mathrm{cal}} \;\geq\; \frac{K\,\ln(2K/\beta)}{2\delta^2}\,\big(1+o(1)\big) \;=\; O\big(K\log K\,/\,\delta^2\big),$$

proving equation 26. □

Remark 11 (Origin and necessity of the $\log K$ factor). The $\ln K$ inside equation 25 enters solely through the requirement that each per-stratum failure probability be at most β/K. Under the *simultaneous-attainment* semantics (balanced strata of equal size n_b, independent calibration subsamples across strata, and the requirement that *all* K strata attain slack δ at constant overall confidence) this logarithm is necessary, not an artifact of the union bound. Formally, if each stratum fails independently with probability q, then the probability that all strata succeed is exactly

$$(1-q)^K \;\leq\; e^{-Kq},$$

so constant overall confidence $1-\beta$ forces $Kq \leq \ln\big(1/(1-\beta)\big)$, i.e. $q = O(1/K)$. The per-stratum failure probability is the Beta lower tail of Step 1, $q = \mathbb{Q}\big(U \leq 1-\alpha-\delta\big)$ with $U \sim \mathrm{Beta}(k_b, n_b+1-k_b)$; the Hoeffding exponent used above matches the binomial rate function up to constant factors in the moderate-deviation regime: at displacement δ, $D\big(c_b-\delta\,\|\,c_b\big) = \delta^2/\big(2c_b(1-c_b)\big) + O(\delta^3) = \Theta(\delta^2)$ (Pinsker's inequality gives the lower bound $D \geq 2\delta^2$), so $q \approx \exp\big(-\Theta\big((n_b+1)\delta^2\big)\big)$ up to sub-exponential factors, and $q = O(1/K)$ forces $n_b = \Omega\big(\ln(K)/\delta^2\big)$. The heuristic step in this necessity argument is only the moderate-deviation equivalence of the per-stratum tail; the product formula $(1-q)^K$ is exact under independence. The statement is restricted to the simultaneous-attainment semantics: weaker semantics (e.g. bounding the *expected* number of violating strata) can evade the $\log K$ factor, as can allocations that correlate the strata.

Remark 12 (Small-n_b strata). If $(n_b+1)\alpha < 1$ then $k_b = n_b+1$ and $\hat{s}_b = +\infty$: equation 25 is vacuous for that stratum while the average guarantee of Theorem 5 still holds. The certificate Cert_b of Theorem 4 remains computable at any $n_b \geq 2$ with $O(n_b^{-1/2})$ concentration, and is exactly the diagnostic instrument for the regime in which quantile recalibration ($n_b \gtrsim \delta^{-2}\ln(K/\beta)$) is not yet feasible.

O Statement and Proof of Theorem 7

We restate the theorem in full.

Theorem 7 (Bound-optimal stratification)**.** *Suppose the stratification variable $\zeta(x)$ takes values in an interval of length Δ, that λ is L-Lipschitz in ζ, and that strata are equal-width in ζ with balanced counts $n_b = n_{\mathrm{cal}}/K$. Then within-stratum dispersion obeys $|\epsilon_b(x)| \le L\Delta/(2K)$, and combining Theorem 2 with Theorem 6, the pointwise coverage slack obeys*

$$\ell(K, n_{\mathrm{cal}}) \;\le\; \underbrace{2\,C_f\,u\left(e^{L\Delta/2K}-1\right)}_{\textit{dispersion (bias)}} \;+\; \underbrace{\sqrt{\tfrac{K\,\ln(2K/\beta)}{2\,n_{\mathrm{cal}}}}+\tfrac{K}{n_{\mathrm{cal}}}}_{\textit{quantification (variance)}} \;\approx\; \frac{C_f\,u\,L\,\Delta}{K}+\sqrt{\frac{K\,\ln(2K/\beta)}{2\,n_{\mathrm{cal}}}}. \tag{27}$$

Minimizing over K gives

$$K^* \;=\; \Theta\Big(\big(n_{\mathrm{cal}}/\ln n_{\mathrm{cal}}\big)^{1/3}\Big), \qquad \min_K \ell(K, n_{\mathrm{cal}}) \;=\; \Theta\Big(\big(\ln n_{\mathrm{cal}}/n_{\mathrm{cal}}\big)^{1/3}\Big). \tag{28}$$

We first bound the within-stratum dispersion under the Lipschitz model, then assemble the trade-off and optimize it.

Lemma 7 (Within-stratum dispersion under equal-width Lipschitz stratification)**.** *Suppose the stratification variable $\zeta(x)$ takes values in an interval of length Δ, that λ is L-Lipschitz in ζ, and that the K strata are equal-width bins of ζ. Then the midpoint m_b of the range of λ over stratum b satisfies*

$$\sup_{x\in S_b}\big|\lambda(x)-m_b\big| \;\le\; \frac{L\Delta}{2K}, \qquad \textit{and consequently} \qquad \big|e^{-\epsilon_b(x)}-1\big| \;\le\; e^{L\Delta/2K}-1 \quad \textit{for } \epsilon_b(x) := \lambda(x)-m_b.$$

Proof. Equal-width bins of a length-Δ interval have width Δ/K, so for $x, x' \in S_b$ the Lipschitz condition gives $\big|\lambda(x)-\lambda(x')\big| \le L\,\big|\zeta(x)-\zeta(x')\big| \le L\Delta/K$; the range of λ over S_b is therefore at most $L\Delta/K$, and every point of the range lies within half the range of its midpoint m_b, giving the first display with $L\Delta/(2K)$. For the consequence,

$$\big|e^{-\epsilon}-1\big| \;=\; \Big|\int_0^{-\epsilon} e^s\,ds\Big| \;\le\; \int_{-|\epsilon|}^{0} e^s\,ds \;=\; 1-e^{-|\epsilon|} \;\le\; e^{|\epsilon|}-1,$$

where the last step uses $e^a-1-\big(1-e^{-a}\big) = \big(e^a-1\big)^2 e^{-a} \ge 0$ for $a \ge 0$; then $|\epsilon_b(x)| \le L\Delta/(2K)$ with monotonicity of $a \mapsto e^a-1$ concludes. □

Proof of Theorem 7. Step 1: assembling the bound equation 27. Apply Theorem 2 to the stratum-b subpopulation with reference level equal to the midpoint m_b of Lemma 7 (the theorem allows any reference level) and Theorem 6 for the marginal relaxation: with probability at least $1-\beta$ over the calibration set, for every x with $b(x) = b$,

$$\ell(x) \;:=\; 1-\alpha-\mathrm{Cov}_{\mathrm{real}}(\hat{s}_b \mid x) \;\le\; C_f\,\tilde{t}_b\left(\big|e^{-\epsilon_b(x)}-1\big|+\mathbb{E}_b\big|e^{-\epsilon_b(X')}-1\big|\right) \;+\; \epsilon_{n_b},$$

where $\tilde{t}_b := c\,\hat{s}_b\,e^{-m_b}$. By Lemma 7, each dispersion term is at most $e^{L\Delta/2K}-1$ (the expectation term by Jensen's inequality for the bound $|e^{-\epsilon_b(X')}-1| \le e^{L\Delta/2K}-1$ applied inside the expectation), so, defining $\ell(K, n_{\mathrm{cal}}) := \sup_x \ell(x)$ on the good event,

$$\ell(K, n_{\mathrm{cal}}) \;\le\; 2\,C_f\,\tilde{t}\left(e^{L\Delta/2K}-1\right) \;+\; \epsilon_{n_b}, \qquad \tilde{t} := \max_b \tilde{t}_b.$$

Two bookkeeping steps convert this into equation 27. First, the quantification term: with $n_b = n_{\mathrm{cal}}/K$,

$$\epsilon_{n_b} \;=\; \sqrt{\frac{\ln(2K/\beta)}{2\,(n_{\mathrm{cal}}/K+1)}} \;\le\; \sqrt{\frac{K\,\ln(2K/\beta)}{2\,n_{\mathrm{cal}}}},$$

and in the relaxed convention of Theorem 2 (Lemma 6), ϵ_{n_b} equals the displayed square-root term plus $O(1/n_b) = O(K/n_{\mathrm{cal}})$, which is the origin of the K/n_{cal} term in equation 27. Second, the constant: under Assumption 5, Step 3 of the proof of Theorem 2 applied per stratum gives $\tilde{t}_b = u\,\big(1+O(\mathbb{E}_b|\epsilon_b| + \epsilon_{n_b})\big)$ with implicit constant $(C_f/f_{\min})\,e^{O(L\Delta)}$ (note $\|\epsilon_b\|_\infty \le L\Delta/2$ uniformly in $K \ge 1$), so $\tilde{t} = u\,(1+o(1))$ along any sequence with $K \to \infty$ and $n_{\mathrm{cal}}/K \to \infty$, and replacing $\tilde{t}$ by u multiplies the dispersion term by $1+o(1)$, a factor absorbed by the Θ-statements below. The displayed $\approx$ form of equation 27 uses $e^{L\Delta/2K}-1 \le \frac{L\Delta}{2K}\,e^{L\Delta/2K} = \Theta(L\Delta/K)$ for $K \to \infty$.

Step 2: optimization. Minimize the two leading terms of equation 27, $\phi(K) := A/K + \sqrt{B(K)}$ with $A := C_f\, u\, L\, \Delta$ and $B(K) := K\, \ln(2K/\beta)/(2n_{\mathrm{cal}})$. Note $\phi(K) \to \infty$ as $K \downarrow 0$ and as $K \to \infty$, so an interior minimizer exists. Differentiate:

$$\phi'(K) \;=\; -\frac{A}{K^2} \;+\; \frac{1}{2}\sqrt{\frac{\ln(2K/\beta)}{2\,n_{\mathrm{cal}}\,K}}\Big(1 + \frac{1}{\ln(2K/\beta)}\Big) \;=\; -\frac{A}{K^2} \;+\; \frac{1}{2}\sqrt{\frac{\ln(2K/\beta)}{2\,n_{\mathrm{cal}}}}\; K^{-1/2}\,\big(1 + o(1)\big).$$

Setting $\phi'(K) = 0$ gives the stationarity condition

$$K^{3/2} \;=\; 2A\sqrt{\frac{2\,n_{\mathrm{cal}}}{\ln(2K/\beta)}}\,\big(1 + o(1)\big),$$

an implicit equation since K appears under the logarithm. Taking logarithms,

$$\tfrac{3}{2}\ln K \;=\; \tfrac{1}{2}\ln n_{\mathrm{cal}} \;-\; \tfrac{1}{2}\ln\ln(2K/\beta) \;+\; O(1).$$

The substitution $\ln K = \frac{1}{3}\ln n_{\mathrm{cal}}$ is self-consistent: it gives $\ln\ln(2K/\beta) = \ln\big(\frac{1}{3}\ln n_{\mathrm{cal}} + O(1)\big) = o(\ln n_{\mathrm{cal}})$, so the fixed point satisfies $\ln K^* = \frac{1}{3}\ln n_{\mathrm{cal}}\,\big(1 + o(1)\big)$ and

$$K^* \;=\; \Theta\Big(\big(n_{\mathrm{cal}}/\ln n_{\mathrm{cal}}\big)^{1/3}\Big).$$

Substituting back, the two leading terms are of equal order:

$$\frac{A}{K^*} \;=\; \Theta\Big(\big(\ln n_{\mathrm{cal}}/n_{\mathrm{cal}}\big)^{1/3}\Big),$$
$$\sqrt{\frac{K^*\,\ln(2K^*/\beta)}{2\,n_{\mathrm{cal}}}} \;=\; \Theta\Big(\Big(\tfrac{(n_{\mathrm{cal}}/\ln n_{\mathrm{cal}})^{1/3}\,\ln n_{\mathrm{cal}}}{n_{\mathrm{cal}}}\Big)^{1/2}\Big) \;=\; \Theta\Big(\big(\ln n_{\mathrm{cal}}/n_{\mathrm{cal}}\big)^{1/3}\Big),$$

and the neglected terms are of smaller order there: $K^*/n_{\mathrm{cal}} = \Theta\big(n_{\mathrm{cal}}^{-2/3}(\ln n_{\mathrm{cal}})^{-1/3}\big) = o\big((\ln n_{\mathrm{cal}}/n_{\mathrm{cal}})^{1/3}\big)$, and the $1 + o(1)$ factor on the dispersion term contributes a relative $o(1)$ correction since $\mathbb{E}_b|\epsilon_b| \le L\Delta/(2K^*) \to 0$ and $\epsilon_{n_{K^*}} \to 0$. Hence $\min_K \ell(K, n_{\mathrm{cal}}) = \Theta\big((\ln n_{\mathrm{cal}}/n_{\mathrm{cal}})^{1/3}\big)$, proving equation 28.

Step 3: general dispersion law. Under the general within-stratum dispersion law $\sigma_b = O(K^{-\gamma})$, $\gamma > 0$ (Step 1 then yields a dispersion term of order $\Theta(K^{-\gamma})$: with reference level m_b, $|\epsilon_b(x)| = O(K^{-\gamma})$ and $|e^{-\epsilon_b(x)} - 1| = \Theta(K^{-\gamma})$ for worst-case strata), the leading bound is $\phi_\gamma(K) = A_\gamma\, K^{-\gamma} + \sqrt{B(K)}$. Stationarity:

$$-\gamma A_\gamma\, K^{-\gamma-1} \;+\; \frac{1}{2}\sqrt{\frac{\ln(2K/\beta)}{2\,n_{\mathrm{cal}}}}\; K^{-1/2}\,\big(1 + o(1)\big) \;=\; 0 \quad\Longrightarrow\quad K^{\gamma+1/2} \;=\; \Theta\Big(\sqrt{n_{\mathrm{cal}}/\ln n_{\mathrm{cal}}}\Big),$$

where the self-consistency $\ln K^* = \frac{1}{2\gamma+1}\ln n_{\mathrm{cal}}\,(1 + o(1))$ follows by the same logarithmic substitution as Step 2. Hence

$$K^* \;=\; \Theta\Big(\big(n_{\mathrm{cal}}/\ln n_{\mathrm{cal}}\big)^{1/(2\gamma+1)}\Big), \qquad \min_K \ell(K, n_{\mathrm{cal}}) \;=\; \Theta\Big(\big(\ln n_{\mathrm{cal}}/n_{\mathrm{cal}}\big)^{\gamma/(2\gamma+1)}\Big),$$

the second display because both leading terms evaluate to $\Theta\big((\ln n_{\mathrm{cal}}/n_{\mathrm{cal}})^{\gamma/(2\gamma+1)}\big)$ at K^*: for the bias term directly, and for the variance term since

$$\Big(\frac{K^*\ln K^*}{n_{\mathrm{cal}}}\Big)^{1/2} \;=\; \Big(\frac{(n_{\mathrm{cal}}/\ln n_{\mathrm{cal}})^{1/(2\gamma+1)}\,\ln n_{\mathrm{cal}}}{n_{\mathrm{cal}}}\Big)^{1/2} \;=\; \Big(\frac{\ln n_{\mathrm{cal}}}{n_{\mathrm{cal}}}\Big)^{\gamma/(2\gamma+1)}.$$

The case $\gamma = 1$ recovers Step 2. □

Remark 13 (Status of the optimality statement). As stated in the main text, equation 28 is the optimal order *of the constructed upper bound* equation 27; no matching minimax lower bound is claimed. Its content is comparative: at $K = 1$ the dispersion term of equation 27 can be $\Theta(1)$ in the Lipschitz worst-case model (whole-space scale dispersion), while the optimized bound is $\widetilde{O}(n_{\mathrm{cal}}^{-1/3})$, with the quantification noise at K^* remaining $\Theta\big((\ln n_{\mathrm{cal}}/n_{\mathrm{cal}})^{1/3}\big)$ of the same order as the bias.

P Statement and Proof of Theorem 8

We restate the theorem in full.

Theorem 8 (End-to-end pointwise guarantee)**.** *Let μ_b denote the stratum mean of $\lambda - \eta$ and define the* effective dispersion

$$\delta_b(x) \;:=\; \lambda(x) - \eta(x) - \mu_b. \tag{29}$$

Under Assumptions 1 and 2, with stratified scalar calibration $\hat{s}_b$, with probability at least $1-\beta$ over the calibration set, simultaneously for every $x \in \mathcal{X}$ (with $b = b(x)$),

$$\mathrm{Cov}_{\mathrm{real}}(\hat{s}_b \mid x) \;\geq\; 1 - \alpha \;-\; \underbrace{C_f\,\tilde{t}_b\,\Big(\,\big|e^{-\delta_b(x)} - 1\big| + \mathbb{E}_b\big[\,\big|e^{-\delta_b(X')} - 1\big|\,\big]\Big)}_{\textit{simulator-shift term (within-stratum effective dispersion, incl. Stage I)}} \;-\; \underbrace{\epsilon_{n_b}}_{\textit{finite-calibration term}}, \tag{30}$$

where $\tilde{t}_b := c\,\hat{s}_b\,e^{-\mu_b}$ and $\epsilon_{n_b} = \sqrt{\ln(2K/\beta)/\big(2(n_b+1)\big)}$ as in Theorem 6. Also, under Assumption 5, $\tilde{t}_b = u\cdot\big(1 + O\big(\mathbb{E}_b|\delta_b| + \epsilon_{n_b}\big)\big)$, so the constant $C_f\,\tilde{t}_b$ is interpretable as $C_f\,u\,(1+o(1))$ whenever the effective dispersion is small.

Remark 14 (Stage-I error must enter pointwise). A global mean term $C_f\,\tilde{t}\,\mathbb{E}\big|e^{\eta(X)} - 1\big|$ cannot control Stage-I error: with $Z \sim \mathrm{Unif}[0,1]$, $\lambda \equiv 0$, and $\eta = 1$ on a set of probability 10^{-3}, pointwise coverage on that set drops from 0.9 to 0.33 while the mean term is of order 10^{-3}. The effective dispersion equation 29 is the correct treatment: it keeps $\eta(x)$ where it acts, at the point x.

Proof of Theorem 8. We use Assumptions 1 and 2 only; in particular, the sub-Gaussian dispersion Assumption 3 does *not* enter this proof (it is used solely by the certificate of Theorem 4), and Assumption 5 is invoked only for the closing order statement. The Stage-I error $\eta(x) = \log q(x) - \log\big(c\,\sigma_{\mathrm{sim}}(x)\big)$ of Assumption 6 is carried pointwise throughout.

Step 1: identity with the effective scale ratio. Since $q(x) = c\,\sigma_{\mathrm{sim}}(x)\,e^{\eta(x)}$, the normalized score under $\mathbb{Q}$ factors as

$$V \;=\; \frac{\sigma_{\mathrm{real}}(X)\,Z}{q(X)} \;=\; e^{\lambda(X)-\eta(X)}\cdot\frac{Z}{c},$$

and since $Z \perp X$, the key identity equation 4 generalizes to

$$\mathrm{Cov}_{\mathrm{real}}(s \mid X = x) \;=\; F_Z\big(c\,s\,e^{-\lambda(x)+\eta(x)}\big) \;=\; F_Z\big(c\,s\,e^{-\mu_b}\,e^{-\delta_b(x)}\big), \qquad b = b(x),$$

with the stratum level μ_b (the stratum mean of $\lambda - \eta$) and the effective dispersion $\delta_b(x) = \lambda(x) - \eta(x) - \mu_b$ of equation 29. The Stage-I pointwise error is thereby absorbed into the effective dispersion alongside the scale-ratio heterogeneity; this absorption is what keeps the bound valid pointwise, since a global mean treatment of η is violated outright (Remark 14).

Step 2: Theorem 2 on the stratum subpopulation. Fix a stratum b and apply Steps 1–2 of the proof of Theorem 2 to the subpopulation $\mathbb{Q}(\cdot \mid X \in S_b)$, with reference level μ_b (the theorem permits any reference level), dispersion field δ_b, threshold $\hat{s}_b$, and $\tilde{t}_b := c\,\hat{s}_b\,e^{-\mu_b}$. The argument uses only the identity of Step 1, the Lipschitz bound $|F_Z(a) - F_Z(b)| \leq C_f|a-b|$ (Assumption 2), and Jensen's inequality for $|\cdot|$; hence, conditionally on the calibration set, for every x with $b(x) = b$,

$$\mathrm{Cov}_{\mathrm{real}}(\hat{s}_b \mid x) \;\geq\; G_b(\hat{s}_b) \;-\; C_f\,\tilde{t}_b\,\Big(\big|e^{-\delta_b(x)} - 1\big| + \mathbb{E}_b\big[\,\big|e^{-\delta_b(X')} - 1\big|\,\big]\Big),$$

where $G_b(t) = \mathbb{Q}(V \leq t \mid X \in S_b)$ and $\mathbb{E}_b$ denotes expectation under $\mathbb{Q}(\cdot \mid X \in S_b)$.

Step 3: marginal relaxation. Theorem 6 supplies, with probability at least $1-\beta$ over the calibration set, $G_b(\hat{s}_b) \geq 1 - \alpha - \epsilon_{n_b}$ simultaneously for all $b = 1, \ldots, K$, with $\epsilon_{n_b} = \sqrt{\ln(2K/\beta)/\big(2(n_b+1)\big)}$; its applicability conditions are met here (G_b continuous under Assumption 2, within-stratum subsamples i.i.d.), and the union bound over strata is already inside the $\ln(2K/\beta)$ factor. Substitution into Step 2 yields equation 30, simultaneously for every x, on the same probability-$(1-\beta)$ event.

Step 4: order of $\tilde{t}_b$ under Assumption 5. Step 3 of the proof of Theorem 2 applies verbatim on each stratum subpopulation (function-value proximity with $\delta_1^{(b)} := C_f\,u\,\mathbb{E}_b|e^{-\delta_b} - 1|$, local strong monotonicity $H_b'(t) \geq f_{\min}e^{-\|\delta_b\|_\infty}$ subject to the same small-dispersion locality regime equation 15, and transport of the two-sided estimation error $\epsilon_{n_b} + 1/(n_b+1)$), giving

$$\tilde{t}_b \;=\; u\cdot\Big(1 + O\big(\mathbb{E}_b|\delta_b| + \epsilon_{n_b}\big)\Big),$$

with implicit constant $(C_f/f_{\min})\, e^{O(\|\delta_b\|_\infty)}$, so $C_f\, \tilde{t}_b = C_f\, u\,(1 + o(1))$ whenever the effective dispersion is small. The bound equation 30 itself, like that of Theorem 2, does not invoke Assumption 5.

Step 5: optional separation of scale-ratio dispersion from Stage-I error. If a separate display of the two sources is desired, the pointwise inequality

$$\left|e^{-\epsilon_b(x)-\eta(x)} - 1\right| \;\le\; \left|e^{-\epsilon_b(x)} - 1\right| \;+\; e^{-\epsilon_b(x)}\left(e^{|\eta(x)|} - 1\right)$$

(with $\epsilon_b(x) := \lambda(x) - \mu'_b$ the scale-ratio part around its own stratum reference μ'_b) follows from $e^{-\epsilon-\eta} - 1 = \left(e^{-\epsilon} - 1\right) + e^{-\epsilon}\left(e^{-\eta} - 1\right)$ and $\left|e^{-\eta} - 1\right| \le e^{|\eta|} - 1$ (the chain proved in Lemma 7); the second term is the *pointwise* Stage-I penalty $e^{|\eta(x)|} - 1$, not a global mean term. □

Remark 15 (The degradation spectrum, assembled). Theorem 8 is the sufficient-data end of the spectrum indexed by the per-stratum real sample size, and the pieces proved above cover every point of it. *(i)* $n_b = 0$*:* no certificate (Cert_b needs $n_b \ge 2$) and no calibration; the stratum falls back to the zero-shot interval s_P with the prior bound of Theorem 1, borrows a neighboring stratum's scalar under an explicit cross-stratum extrapolation assumption with an additional error term, or is flagged as requiring real data. *(ii)* $2 \le n_b \lesssim \delta^{-2} \ln K$*:* Cert_b diagnoses the stratum's zero-shot degradation (Theorem 4) and Theorem 5's average guarantee applies, while realized per-stratum guarantees are vacuous (Remark following Theorem 6). *(iii) Sufficient* n_b*:* equation 30 gives pointwise control, whose two terms are the within-stratum effective dispersion (scale-ratio level absorbed by $\hat{s}_b$ (Corollary 2), dispersion localized by stratification (Theorem 7), Stage-I error absorbed alongside) and the finite-calibration noise $O\big(\sqrt{K \ln K \,/\, n_{\mathrm{cal}}}\big)$, driven to $\widetilde{O}(n_{\mathrm{cal}}^{-1/3})$ by the optimal K^*. When the deployment protocol runs audit-then-calibrate, the error budgets compose by union bound: $2K\beta'$ (Theorem 4), γ (Proposition 4), and β (Theorems 6 and 8), with the deterministic simulator Monte Carlo and Stage-I training errors entering through the δ-terms and the effective dispersion, respectively.