NOT ALL MISLABELS ARE EQUAL: EFFICIENT TASK-AWARE AUDITING WITH CONFIDENCE SEQUENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study efficient, adaptive auditing of categorical labels in a fixed dataset, motivated by biodiversity-credit programs in which expert verification of species records is costly and downstream scores (e.g., payment) computed from corrected labels are more sensitive to errors on certain labels than on others. Our framework couples anytime-valid inference with a task-aware adaptive auditing policy, representing label corrections through a confusion matrix and prioritizing which data points to inspect next based on the uncertainty within their class labels and the corresponding influence of these labels on the downstream score function. For linear scores such as the mislabel rate, our priority-aware approach certifies the overall label error rate to ± 2 percentage points with 24% and 11% fewer audits compared to random sampling on biodiversity data from Kenya and a widely-used CIFAR-10N benchmark, respectively. For nonlinear scores such as the Hill–Shannon diversity, we exploit a linear-envelope representation that reduces inferences to a family of linear-score audits; priority-aware sampling again reaches the target precision with 36% fewer audits. When the cost of distribution-free guarantees becomes prohibitive, we develop a model-based Bayesian alternative that trades assumption-free validity for considerably fewer audits. Together, these results show task-aware auditing can significantly reduce the number of expert audits needed while providing finite-sample, anytime-valid guarantees.

1 INTRODUCTION

Large labeled datasets are routinely used in consequential inference tasks, from model evaluation to scientific analysis. However, the labels in these datasets are not always correct, with annotation errors documented in settings ranging from healthcare (Shi et al., 2024; Rädsch et al., 2023) to remote sensing (Pelletier et al., 2017). Even widely used machine-learning benchmarks contain nontrivial label noise, with one study estimating an average error rate of 3.3% (Northcutt et al., 2021). Especially in consequential settings, practitioners may want rigorous uncertainty quantification, but doing so requires verifying at least some of the labels. When expert verification of an entire dataset is impractical, two natural questions arise: How many data points must be audited to certify a downstream quantity with a desired level of confidence? Which records should be audited next to reduce that uncertainty as efficiently as possible?

When audits are costly, confidence sequence methods are useful as they provide confidence intervals that are simultaneously valid over all sample sizes, allowing the auditor to inspect data sequentially and stop adaptively without invalidating the coverage guarantee. Recent work develops such guarantees for sampling without replacement (Waudby-Smith & Ramdas, 2020; Ryu & Wornell, 2024) to certify finite-population quantities, e.g., population mean, but these methods were not designed for the precision audits need in practice: on our data of 33,403 labels, the prior–posterior-ratio confidence sequence of Waudby-Smith & Ramdas (2020) needs a median of 4,118 audits to certify the mislabel rate to ± 2 percentage points at 95.5% confidence, against 656 for our method. On unlabeled datasets, active-evaluation methods adaptively select samples to label to estimate quantities such as model risk or predictive performance (Sawade et al., 2012; Kossen et al., 2021; 2022; Mittal et al., 2024), and recent work has begun to combine adaptive label acquisition with anytime-valid inference (Xu et al., 2024). Our setting brings adaptive data acquisition and anytime-valid inference together in a different form: the data are already labeled but may be incorrect, the unknown corrections are categorical and

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.

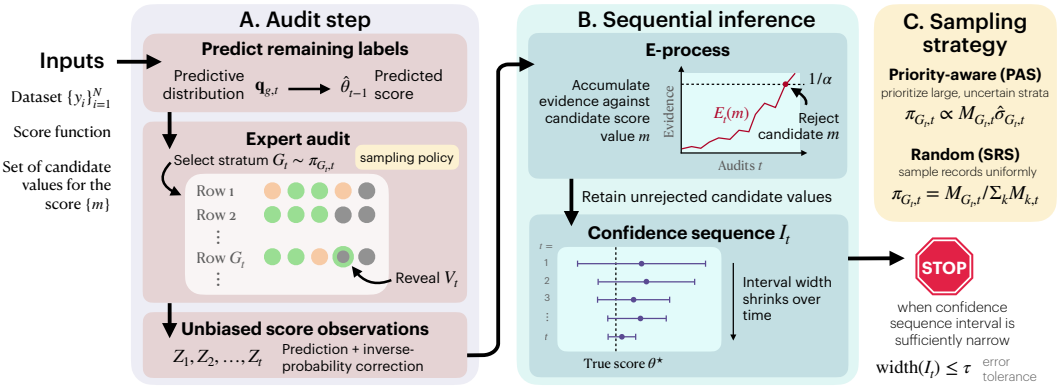

Figure 1: **Anytime-valid, priority-aware auditing framework.** **(A)** Each audit uses past audits to predict the unverified labels and the target score. An expert audit V_t corrects the predicted score through inverse-probability weighting to produce an unbiased score observation Z_t. **(B)** E-processes accumulate evidence against each candidate score value m, reject values with sufficient evidence, and use remaining values to form a confidence sequence I_t, stopping once the width falls below a tolerance τ. **(C)** Our priority-aware sampling strategy directs reviews toward large strata with high predicted score uncertainty. We compare against random sampling as a baseline, but note that different sampling strategies only change the efficacy, not the validity of the confidence sequences.

structured, the downstream target may be nonlinear or non-smooth, and the auditor must adaptively decide which records to verify without sacrificing validity.

Motivation: Auditing for biodiversity credits. Biodiversity credits are purchased by companies and investors to fund nature conservation and restoration worldwide, compensating individual landowners for measurable improvements in ecosystem health (Wunder et al., 2025). The market is estimated to grow from US$8M annually to $69B by 2050 (World Economic Forum, 2023), but credits are burdened by high overhead costs and concerns about trust and verification (Aide, 2024). Each project may generate tens or hundreds of thousands of sensor records documenting wildlife presence and abundance, making full expert audit impractical. Automated species classifiers reduce labeling cost but can have substantial error: one widely used classifier had an 18.2% error rate on an out-of-sample dataset (Tabak et al., 2019). Because credit payments depend on these metrics, label errors have direct financial consequences, creating a need for efficient expert auditing. An aggregate mislabel rate (e.g., "13% of labels are incorrect") is insufficient because the direction of an error determines its downstream consequence. Incorrectly labeling cattle as lion, for example, may falsely introduce a rare species and substantially inflate estimated richness and taxonomic diversity (and therefore payment), whereas confusing two common, closely related species may have little effect. An effective audit should therefore account not only for where errors are likely, but also for how potential corrections affect the downstream scoring function.

The same structure arises in myriad settings. In model evaluation, for example, noisy benchmark labels impact downstream quantities such as F_1 score, balanced accuracy, or model rankings (Northcutt et al., 2021), each of which may depend nonlinearly on the corrected labels. In epidemiology, label corrections may change prevalence estimates (Haile & Kronthaler, 2025), while in content moderation they may alter policy-compliance rates.

To address these challenges, we make the following contributions:

- **A general framework for efficient label auditing (fig. 1).** We introduce a task-aware auditing framework for fixed categorical datasets that combines anytime-valid uncertainty quantification with adaptive expert review. The framework certifies downstream quantities while directing audits toward the label corrections that matter most, reducing the number of costly expert audits needed in practice.
- **Efficient anytime-valid inference for structured nonlinear scores.** For nonlinear scores such as Hill–Shannon diversity, we exploit linear-envelope representations to build score-specific confidence sequences from our linear auditing framework. The resulting confidence

sequences retain finite-sample, anytime-valid guarantees while supporting efficient priority-aware auditing.

- **A model-based (Bayesian) extension within the same framework.** When distribution-free guarantees require an impractically large number of audits, we retain the same task-aware strategy but use a Bayesian model of the confusion matrix for uncertainty quantification, trading assumption-free validity for greater statistical efficiency.

2 Problem Formulation

A provided dataset contains N records. Record $i \in [N]$ has a submitted label y_i (known) and verified label v_i (unknown until verification). Let $\mathcal{L} := \{1, \dots, L\}$ be the full normalized label vocabulary, and let $\mathcal{R} := \{y_i : i \in [N]\} \subseteq \mathcal{L}$ be the set of submitted labels with $R := |\mathcal{R}|$ elements. Thus $y_i \in \mathcal{R}$ and $v_i \in \mathcal{L}$.

Confusion matrix. A scalar mislabel rate records only how often $y_i \neq v_i$ across the entire dataset; it does not capture error rates across different classes. We therefore track a full confusion matrix, which specifies the number of mislabels between any two pairs of labels. The unknown finite-population confusion matrix is $\mathbf{C}^\star \in \mathbb{Z}_{\geq 0}^{R \times L}$, where each entry $C_{rs} := \#\{i : y_i = r,\ v_i = s\}$ indicates the number of data points submitted with row r and verified with column s. Row r contains records submitted as label r, while column s contains records verified as label s. The row sums $N_r := \sum_{s \in \mathcal{L}} C^\star_{rs} = \#\{i : y_i = r\}$ are known before auditing.

Score functions. We use *score function* to denote any downstream quantity computed from the corrected labels. Formally, a score function maps the submitted-to-verified confusion matrix $\mathbf{C}$ to a scalar outcome of interest, such as an error rate, model-evaluation metric, or biodiversity score. One important property is that the score function is used for finite-population inference. Unlike conventional superpopulation inference, when auditing labels we are only interested in the particular submitted dataset rather than by an i.i.d. sample from an underlying distribution. That is, we only care about sample statistics, not population statistics. A second property is that score functions can be assessed by the verified labels, i.e., the score function depends on $\mathbf{C}$ only through its corrected column totals which we write as the corrected composition $p(\mathbf{C}) \in \Delta^{L-1}, p_s(\mathbf{C}) = \sum_r C_{rs}/N$, with Δ^{L-1} the probability simplex over $\mathcal{L}$; we abbreviate $p^\star = p(\mathbf{C}^\star)$. For the biodiversity credits setting, the score function is challenging as it is nonlinear and nonsmooth; details in appendix A.

Strata: splitting rows by metadata. Records often carry metadata, such as the source, location, device, or time period, that may be associated with different correction patterns. Splitting each row by these metadata creates smaller groups of records which we call *strata*. We write $g = (r, a)$ for the stratum of submitted label r and metadata value a. The size of each stratum is known before any audit, and each row is the union of its strata. When no metadata is used, each row is its own stratum. Stratification is useful when correction behavior varies across metadata-defined subgroups, because it allows the audit to target uncertainty more precisely. In our biodiversity application, a refers to a camera station, where corrections are strongly clustered.

3 Confidence Sequences for Auditing Linear Score Functions

We first explore how to turn expert audits into a *confidence sequence*, illustrated in fig. 1b. Confidence sequences provide uncertainty intervals that are valid after every audit, allowing the auditor to monitor uncertainty and stop as soon as desired precision is reached. For linear score functions, we show that each randomized audit can be transformed into an unbiased, prediction-corrected estimate of the score for the full dataset. Thus the confidence sequence remains valid for any predictable audit policy, while better sampling choices reduce its variance and therefore the number of expert audits required.

Let $\mathbf{D} \in \mathbb{R}^{R \times L}$ be a deterministic contribution matrix, where the (r, s) element $D_{rs} \in [0, 1]$ is the contribution to the score from a data point submitted as label r and verified as label s and is known in advance. For example, when the score function is the mislabel rate, $\mathbf{D}$ is a binary matrix with $D_{r,s} = \mathbb{1}\{r \neq s\}$. We denote the r-th row of this matrix as $\mathbf{d}_r^\top$.

We begin with linear scores, for which each record contributes a fixed, known amount determined by its submitted and verified labels. A score is linear if it can be expressed as $\theta^\star = \frac{1}{N}\sum_{i=1}^{N} D_{y_i,v_i} = \frac{1}{N}\sum_{r\in\mathcal{R}}\sum_{s\in\mathcal{L}} C^\star_{rs} D_{rs}$. If all verified labels were known, $\theta^\star$ could be computed exactly. During the audit, however, only a subset of the data have been reviewed.

Constructing an unbiased audit observation. We first construct a quantity Z_t whose conditional mean, given everything seen so far, is the target score $\theta^\star$:

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \theta^\star \qquad \text{for every labeling.} \tag{1}$$

Audit t examines record $J_t \in [N]$, which lies in the drawn stratum G_t and has verified label $V_t = v_{I_t}$. Let $\mathcal{A}_{t-1} = \{J_1, \ldots, J_{t-1}\}$ be the records reviewed before audit t, $\mathcal{U}_{t-1} = [N] \setminus \mathcal{A}_{t-1}$ the unreviewed ones, and $\mathcal{U}_{g,t-1}$ those of stratum g, with sizes $M_t = |\mathcal{U}_{t-1}| = N - t + 1$ and $M_{g,t} = |\mathcal{U}_{g,t-1}|$. The audit history $\mathcal{F}_{t-1}$ consists of $\{(G_k, J_k, V_k)\}_{k=1}^{t-1}$. For unreviewed records in stratum g, we express the mean as $\mu_{g,t} = \frac{1}{M_{g,t}}\sum_{i\in\mathcal{U}_{g,t-1}} D_{y_i v_i}$ and variance as $\sigma^2_{g,t} = \frac{1}{M_{g,t}}\sum_{i\in\mathcal{U}_{g,t-1}}(D_{y_i,v_i} - \mu_{g,t})^2$. The linear score can be split into:

$$N\theta^\star = \sum_{i\in\mathcal{A}_{t-1}} D_{y_i v_i} + \sum_{i\in\mathcal{U}_{t-1}} D_{y_i v_i} = \sum_{i\in\mathcal{A}_{t-1}} D_{y_i v_i} + \sum_{g} M_{g,t}\,\mu_{g,t}. \tag{2}$$

After $t-1$ audits, the first sum is known and the second must be learned (or estimated) by the auditor.

Before audit t, using only $\mathcal{F}_{t-1}$, the auditor fixes a selection probability $\pi_{g,t}$ and a predicted verified-label distribution $\mathbf{q}_{g,t} \in \Delta^{L-1}$ (from among the $M_{g,t}$ remaining, unreviewed records) for each stratum g. It draws a stratum G_t with probability $\pi_{G_t,t}$ and audits a single record I_t chosen uniformly from $\mathcal{U}_{G_t,t-1}$, so each unreviewed record of stratum g is audited next with probability $\pi_{g,t}/M_{g,t}$.

For audit t, predict θ via the estimator:

$$\hat{\theta}_{t-1} = \frac{1}{N}\cdot\Big(\sum_{i\in\mathcal{A}_{t-1}} D_{y_i,v_i} + \sum_{g} M_{g,t}\cdot \mathbf{d}_{r(g)}^\top \mathbf{q}_{g,t}\Big)\,. \tag{3}$$

The next randomized audit corrects the prediction $\hat{\theta}_{t-1}$. Because strata may be sampled nonuniformly, simply plugging in the newly verified contribution would generally lead to a biased estimate. Instead, we use an inverse-probability-weighted correction in the spirit of Horvitz & Thompson (1952). Replacing each $\mu_{g,t}$ by its prediction $\mathbf{d}_{r(g)}^\top \mathbf{q}_{g,t}$ and correcting the resulting residual gives

$$Z_t = \hat{\theta}_{t-1} + \frac{M_{G_t,t}}{N\cdot\pi_{G_t,t}}(D_{r(G_t),V_t} - \mathbf{d}_{r(G_t)}^\top \mathbf{q}_{G_t,t}). \tag{4}$$

The prediction term acts as a control variate, analogous to Särndal et al. (2003). The inverse-probability-weighted correction makes Z_t unbiased for $\theta^\star$ for any predictable predictions and sampling probabilities, as formalized in lemma 1.

Lemma 1 (One audit, one unbiased observation)**.** *If D, $q_{\cdot,t}$ and $\pi_{\cdot,t}$ are predictable and $\pi_{g,t} > 0$ whenever $M_{g,t} > 0$, then $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \theta^\star$ for every labeling, and with $b_{g,t} = \mu_{g,t} - q_{g,t}^\top D_{r(g)\cdot}$,*

$$\mathrm{Var}(Z_t \mid \mathcal{F}_{t-1}) = \frac{1}{N^2}\Big[\sum_g \frac{M_{g,t}^2}{\pi_{g,t}}\big(\sigma^2_{g,t} + b^2_{g,t}\big) - \Big(\sum_g M_{g,t} b_{g,t}\Big)^2\Big]. \tag{5}$$

Prediction quality affects efficiency, but not validity. In experiments, we set $\mathbf{q}_{g,t}$ to the posterior mean of the hierarchical Dirichlet model (Gelman et al., 1995), fitted to the audits observed before time t.

E-processes. Following Waudby-Smith & Ramdas (2024), fix a candidate value m and predictable stakes $\lambda_i^{\pm}(m) \geq 0$. Under the null hypothesis $\theta^\star = m$, lemma 1 makes each factor below mean one:

$$E_t^+(m) = \prod_{i\leq t}\big\{1 + \lambda_i^+(m)(Z_i - m)\big\}, \qquad E_t^-(m) = \prod_{i\leq t}\big\{1 - \lambda_i^-(m)(Z_i - m)\big\}. \tag{6}$$

If $m = \theta^\star$, both processes are nonnegative martingales with initial value one. By Ville's inequality (lemma 4), each crosses $2/\alpha$ with probability at most $\alpha/2$, uniformly over time.

The only requirement on the stakes $\lambda^{\pm}$ is that they are predictable and keep every factor nonnegative, whatever label the expert reveals. This is always achievable because $D_{rs} \in [0, 1]$, so the range of Z_t is known before each audit. Section F.1 gives this range and the stakes used in our experiments.

Theorem 2 (Confidence sequence for a linear score)**.** *Let* $I_t = \{m : E_s^+(m) < 2/\alpha, E_s^-(m) < 2/\alpha \text{ for all } s \le t\}$. *Then* $\Pr\{\theta^\star \in I_t \,\forall t\} \ge 1 - \alpha$ *for every labeling.*

The proof (section F.1) applies Ville's inequality to the two e-processes at $m = \theta^\star$. Ratios of linear scores reduce to the same test: the recall $\rho^\star = C^\star_{ss} / \sum_r C^\star_{rs}$ of species s and the F_1 score are handled the same way (appendix H).

4 Priority-Aware Sampling

The confidence sequence in section 3 determines when an audit may stop. We now consider what to audit next. Because different label corrections can vary greatly in their impact on the downstream score, we aim to prioritize expert audits towards rows whose unresolved labels contribute most to uncertainty in that score, a process illustrated in fig. 1a.

Suppose $b_{g,t} = 0$, i.e., $\mu_{g,t} = \mathbf{d}_{r(g)}^\top \mathbf{q}_{g,t}$. The variance $\frac{M_{g,t}^2}{\pi_{g,t}}(\sigma_{g,t}^2 + b_{g,t}^2)$ is minimized over probability vectors π s.t. $\pi_{g,t} \propto M_{g,t}\sigma_{g,t}$. This inspires us to review a stratum/row in proportion to its size times its uncertainty.

Proposition 3 (Priority-aware allocation)**.** *If the predictions are exact,* (5) *is minimized subject to floors* $\pi_{g,t} \ge \ell_{g,t}$ *by* $\pi_{g,t} = \max\{\ell_{g,t}, M_{g,t}\sigma_{g,t}/\nu\}$*, with* ν *normalizing. Without floors the minimum is* $N^{-2}(\sum_g M_{g,t}\sigma_{g,t})^2$*, the Neyman variance (Neyman, 1934).*

$M_{g,t}$ is the number of remaining observations to be audited in stratum g and can be calculated from known values. We can estimate the standard deviation $\sigma_{g,t}$ using a heuristic estimator. When the scoring function is the mislabel rate, the contribution is Bernoulli so if $\hat{e}_{g,t} = \mathbf{d}_{r(g)}^\top \mathbf{q}_{g,t} = \widehat{\Pr}(\text{verified label} \neq y_g \mid \mathcal{F}_{t-1})$ is the predicted error rate in stratum g, then the estimated standard deviation is $\hat{\sigma}_{g,t} = \sqrt{\hat{e}_{g,t}(1 - \hat{e}_{g,t})}$.

Then to ensure continued exploration, we mix the resulting priority allocation with sampling proportional to the number of remaining records:

$$\pi_{g,t} = \varepsilon \frac{M_{g,t}}{\sum_k M_{k,t}} + (1 - \varepsilon) \frac{M_{g,t}\hat{\sigma}_{g,t}}{\sum_k M_{k,t}\hat{\sigma}_{k,t}} \,. \tag{7}$$

The first term guarantees continued exploration of all remaining records, while the second approximates the variance-minimizing allocation of proposition 3. The end-to-end priority-aware sampling procedure is presented as algorithm 1 in appendix C.

4.1 Experimental Results: Certifying the Mislabel Rate

Table 1 compares five methods for certifying the overall mislabel rate on the Kenyan biodiversity and CIFAR-10N datasets. Our priority-aware sampling (PAS) method and simple random sampling (SRS) use the same confidence sequence approach, in theorem 2, and differ only in the selection of the next record.

On the Kenyan biodiversity dataset, PAS reaches a half-width of two percentage points after a median of 656 expert reviews, compared with 864 under SRS, a reduction of 24%. On CIFAR-10N, PAS reaches the same target after 768 audits, compared with 862 under SRS, a reduction of 11%. The smaller CIFAR gain is consistent with its more balanced submitted-label rows and less concentrated residual uncertainty. In both settings, PAS retains the anytime-valid guarantee of the underlying confidence sequence while reaching the target precision sooner. A more in-depth empirical study of the mechanism driving priority-aware upside versus SRS is in section D.3.

The remaining methods provide established random-sampling benchmarks. The prior–posterior-ratio method (Waudby-Smith & Ramdas, 2020) is anytime-valid but does not exploit priority-aware allocation. The Wald interval is an asymptotic reference computed using the true population variance, while Serfling's inequality provides a finite-sample bound for sampling without replacement. PAS

Table 1: Audits required to certify the mislabel rate to a half-width of 2 percentage points at 95.5% nominal confidence. The Kenyan biodiversity dataset has a true mislabel rate of 17.97% and CIFAR-10N of 17.23%. We report median stopping times over 20 audit paths.

			# audits to ±2 percentage points	
Method	**Sampling strategy**	**Guarantee**	**Kenyan biodiversity**	**CIFAR-10N**
PAS (ours)	priority-aware	finite-sample, anytime	**656**	**768**
SRS	random	finite-sample, anytime	864	862
Wald	random	asymptotic, fixed n	1,481	1,434
PPR	random	finite-sample, anytime	4,118	3,973
Serfling	random	finite-sample, fixed n	4,154	4,332

reaches the two-percentage-point target substantially earlier than all three benchmarks on both datasets. The trajectories of the confidence sequence widths are shown in fig. 2; the qualitatively similar CIFAR-10N comparison appears in section D.2.1. Additional settings for the five procedures are given in section D.1.1 and section D.2.1.

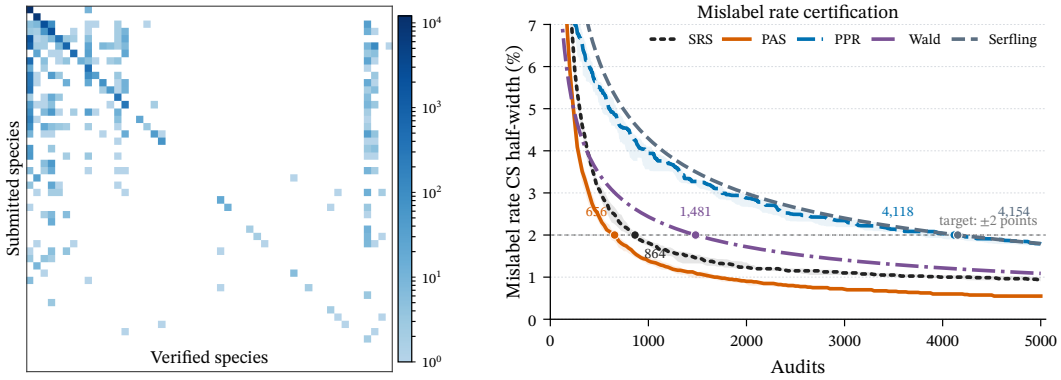

Figure 2: Audits on the Kenyan biodiversity dataset. **(Left)** Submitted-to-verified confusion matrix. Diagonal entries are correct labels, off-diagonal entries are mislabels. Color indicates the number of records on a log scale. **(Right)** Half-width of the 95.5% mislabel-rate interval as the number of audits increases, comparing our PAS approach against four baselines. Curves show medians over 20 audit paths with the 25th–75th percentiles. Dots mark the audits required to reach a half-width of two percentage points. The corresponding CIFAR-10N figure (section D.2.1) exhibits the same relative performances, and numerical results for both settings are reported in table 1.

5 Auditing Nonlinear Scores via Linear Envelopes

In practice, downstream scores are often nonlinear, so the linear construction of section 3 cannot be applied directly with a single fixed contribution matrix. For concave scores, we show they nevertheless admit an anytime-valid audit by exploiting their representation as envelopes of linear scores: a single predictable tangent direction yields a valid upper confidence bound, while the lower bound requires ruling out every plausible label composition, which we certify by convex optimization. For instance, species diversity (part of the biodiversity score function, see appendix A) can be written as $P_2 := \exp H(p(\mathbf{C}^\star))$, where $H(p) = -\sum_{s\in\mathcal{L}} p_s \log p_s$ is the Shannon entropy of the corrected label proportions. We develop this construction for an entropy term and apply the same principle to other nonlinear scores on CIFAR-10N.

Concave scores are envelopes of linear scores. Let H be a concave, differentiable score of the corrected labels. For any reference $\bar{p} \in \Delta^{L-1}$, concavity implies that the tangent plane to H at $\bar{p}$ is an upper bound. Using $\sum_s p_s = 1$, the tangent plane is a linear score,

$$H(p) \;\le\; \sum_{s\in\mathcal{L}} p_s\, \psi_s(\bar{p}), \qquad \text{where } \psi_s(\bar{p}) = H(\bar{p}) + \partial_s H(\bar{p}) - \sum_j \bar{p}_j\, \partial_j H(\bar{p}), \tag{8}$$

with equality at $\bar{p} = p$. Hence $H(p) = \min_{\bar{p}} \sum_s p_s \psi_s(\bar{p})$: a record verified as s contributes the known amount $\psi_s(\bar{p})$ to the linear score indexed by $\bar{p}$. For entropy, $\psi_s(\bar{p}) = \log(1/\bar{p}_s)$, and the tangent plane is the cross-entropy $\mathrm{CE}(p, \bar{p}) := \sum_s p_s \log(1/\bar{p}_s) = H(p) + \mathrm{KL}(p\|\bar{p})$. The inequality in (8) is Gibbs' inequality $\mathrm{KL} \geq 0$ (Cover et al., 1991).

Upper bound: one linear audit. Any predictable reference $\bar{p}$ gives a valid upper confidence bound, its choice affects only the width but not validity. We therefore take the best guess of $p^\star$ available before the audit, i. e., the plug-in composition $\hat{p}_{t-1}$, which counts audited records at their verified labels and unaudited records at their predicted distributions $q_{g,t}$. Auditing the resulting linear score as in section 3 gives the upper confidence sequence U_t^H (appendix G).

Lower bound: ruling out every composition. A tangent plane lies above H, so one cross-entropy cannot certify a lower bound. To show $H^\star > h$ we must show $\mathrm{CE}(p^\star, p) > h$ for *every* composition p still consistent with the audits, $\mathcal{P}_t = \{p \in \Delta^{L-1} : p_s \geq O_{t,s}/N\}$, where $O_{t,s}$ counts audited records verified as s after t audits. Following Manole & Ramdas (2023), we form for each candidate p an e-process $E_t^L(p; h)$ whose factors have conditional mean at most one at $(p^\star, H^\star)$, using a single predictable stake shared by all candidates, and rule out h once $E_t^L(p; h) \geq 1/\alpha_-$ for all $p \in \mathcal{P}_t$, where α_- is the error budget of the lower end. Sharing one stake and one prediction correction across all candidates makes the logarithm of each e-process convex in p, so this minimum is certified by convex optimization rather than enumeration (appendix G). The total budget $\alpha = \alpha_+ + \alpha_-$ is split between the two ends; together with U_t^H, the interval $[L_t^H, U_t^H]$ is the desired confidence sequence. The end-to-end procedure is presented as algorithm 2 in appendix C.

5.1 Experimental Results: Certifying Nonlinear Scores

On the Kenyan biodiversity dataset, PAS reaches the prespecified 5% relative accuracy target for pooled Hill–Shannon diversity P_2 after a median of 1,625 audits, compared with 2,550 under SRS (fig. 3). This saves 925 audits, or 36% of the SRS audit cost. The advantage grows over the audit: relative to paired SRS, PAS produces confidence sequences that are approximately 26% narrower after 2,000 audits and 30% narrower after 3,000. PAS and SRS use the same 95.5% finite-sample, anytime-valid convex-profile certificate and differ only in their predictable allocation of audits. Every stored interval contained the fixed-population truth.

The CIFAR-10N experiments test a different class of nonlinear scores. We define worst-three precision as the mean precision of the three least precise submitted classes $g_{\mathrm{prec}}(C) = \frac{1}{3} \sum_{r \in \mathcal{B}_3(C)} \frac{C_{rr}}{N_r}$, where $\mathcal{B}_3(C)$ indexes the three smallest classwise precision values C_{rr}/N_r. PAS reaches confidence-sequence width 0.08 after 3,750 audits, whereas SRS requires 6,250. We similarly define worst-three taxonomic risk as $g_{\mathrm{tax}}(C) = \frac{1}{3} \sum_{r \in \mathrm{Top}_3(\tau)} \tau_r$, where $\tau_r = N_r^{-1} \sum_s C_{rs} d_{rs}$ and d_{rs} is the normalized taxonomic distance (computed using WordNet 3.0) between classes r and s. PAS reaches width 0.05 at the same respective audit counts. PAS therefore saves 2,500 audits for each target. At 10,000 audits, its median intervals are 21% narrower for worst-three precision and 19% narrower for worst-three taxonomic risk.

These settings are discussed in greater detail in section D.1.2 and section D.2.2.

6 A Bayesian, Model-Based Alternative for Efficient Auditing

The certificates of sections 3 and 5 make no assumptions about how the labels were generated. This assumption-free validity provides strong robustness, but can come at a substantial cost in statistical efficiency, requiring an impractical number of expert audits. A natural alternative is to introduce a statistical model for how submitted labels are corrected, allowing information to be shared across related labels and strata when inferring the composition of unaudited records. This model-based approach can substantially reduce the number of audits required. The compromise is that the resulting uncertainty is no longer distribution-free, but depends on the adequacy of the model. Importantly, the same task-aware auditing framework accommodates both approaches.

For a submitted label r and stratum g, let θ_r denote the overall correction distribution for label r and $\theta_{r,a}$ be the strata-specific correction distribution. We use a hierarchical Dirichlet–multinomial

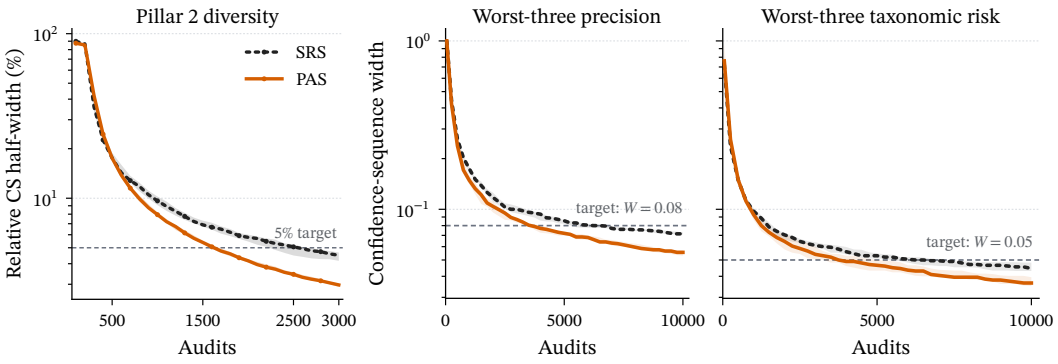

Figure 3: Anytime-valid certification of nonlinear scores under PAS and SRS. **(Left)** Pooled diversity P_2 on the Kenyan biodiversity dataset. **(Middle)** Worst-three precision on CIFAR-10N. **(Right)** Worst-three taxonomic risk on CIFAR-10N. Curves show median confidence-sequence widths with interquartile ranges over 20 paired audit paths for Kenya and 10 for CIFAR-10N.

Table 2: On the Kenya biodiversity data, the number of Bayesian audits required to estimate P_2 to 5% relative accuracy at 95.5% nominal credibility under the model. Results are over 100 audit paths per sampling strategy; all paths reached the target within 1,000 audits. Coverage is the fraction of audit paths whose interval contains the true P_2 at stopping.

	# audits to 5% relative accuracy		
Sampling strategy	**Mean**	**Median**	**Coverage at stop**
Priority-aware (PAS)	**248**	**250**	93%
Random	493	500	89%

model (Kim & Ghahramani, 2012): $\theta_r \sim \text{Dirichlet}\big(\kappa_r\{\beta \mathbf{e}_r + (1-\beta)\bar{\mathbf{g}}\} + \mathbf{A}_r\big)$, and $\theta_{r,a} \mid \theta_r \sim \text{Dirichlet}\big(\kappa_c\,\theta_r + \mathbf{A}_{r,a}\big)$, with unaudited labels in stratum (r,a) drawn from distribution $\text{Multinomial}\big(M_{r,a},\,\theta_{r,a}\big)$. Here, $\mathbf{A}_r$ and $\mathbf{A}_{r,a}$ are counts of the verified labels audited so far, with $M_{r,a}$ unaudited records. Then $\mathbf{e}_r$ places prior mass on the submitted label r being correct, and $\bar{g}$ summarizes the smoothed share of verified labels observed across all labels. The concentration parameters κ_r and κ_c control how strongly information is shared across labels and strata. As audits accumulate within a stratum, its local counts increasingly dominate the shared prior (appendix F).

The model-based audit repeatedly selects and verifies records using the same priority-aware sampling policy (section 4). At audit t, we select a stratum G_t, sample one unverified record uniformly from that stratum, reveal its verified label V_t, and update the observed correction counts. Every 25 audits, we propagate posterior uncertainty to the downstream score by drawing stratum-level correction distributions from the posterior and, conditional on each draw, imputing verified labels for all remaining unaudited records. Each posterior draw defines a complete, corrected dataset consistent with all expert audits observed so far. We compute P_2 on each completed dataset, producing a posterior distribution over the nonlinear downstream score, without relying on a linear approximation. We report the 2.25th and 97.75th percentiles at a 95.5% credible interval and stop once the interval satisfies the prespecified 5% relative-accuracy target.

If the labels were generated by this model, the interval contains the true P_2 with probability 95.5%. That is the key difference from sections 3 and 5, whose guarantees hold without modeling assumptions.

6.1 Experimental Results: Efficient Model-Based Auditing

Our results in table 2 show: (1) With priority-aware selection, the Bayesian approach requires far fewer audits, achieving 5% relative accuracy after a mean of 248 audits — about 6.6× fewer than the distribution-free certificate. (2) Priority-aware selection is as essential in the Bayesian case as it was in the model-free case, halving the cost compared to choosing records at random instead (requires 493 audits). (3) The model is overconfident. The interval covers the truth on 93% of audits, below

the nominal 95.5%, and on some audits it leans low. With random selection coverage falls to 89%. On this dataset, the assumptions that each submitted label is probably right and that errors follow shared patterns, are approximately but not exactly true.

7 Related Work

Anytime-valid inference. Classic finite-population concentration bounds such as Serfling's inequality provide finite-sample guarantees for sampling without replacement at a fixed sample size (Serfling, 1974; Bardenet & Maillard, 2015). Confidence sequences, first introduced by Darling & Robbins (1967), instead provide confidence sets that remain simultaneously valid over time, permitting continuous monitoring and adaptive stopping. Recent work has developed nonparametric, nonasymptotic constructions (Howard et al., 2021; Ramdas et al., 2023). For sampling without replacement from a fixed finite population, Waudby-Smith & Ramdas (2020) develop prior–posterior-ratio, Hoeffding-type, and empirical-Bernstein confidence sequences. The closest works combine anytime-valid inference with adaptive, nonuniform sampling. Shekhar et al. (2023) construct confidence sequences for a finite-population weighted mean under adaptive sampling without replacement, using importance-weighted betting martingales and fixed side information as a control variate. Waudby-Smith et al. (2024) give anytime-valid inference under adaptively chosen sampling probabilities for off-policy evaluation. These methods target a mean of scalar values, and the sampling priority depends on the values themselves (their size, or a prediction of them) rather than on a downstream score. Our section 3 builds on importance-weighted confidence sequences for adaptive finite-population sampling, but exploits the categorical correction structure to construct score-specific prediction corrections and sampling policies. Beyond linear scores, we develop anytime-valid certificates for nonlinear functionals such as entropy using linear-envelope and convex-profile constructions.

Active evaluation. A complementary literature studies how to allocate expensive labels for efficient evaluation. Early active-evaluation methods choose test points to label in order to compare or estimate predictive performance (Sawade et al., 2012; Kossen et al., 2021; 2022) while more recent work formulates adaptive evaluation as sequential planning or active statistical inference (Mittal et al., 2024; Zrnic & Candès, 2024). Li et al. (2025) develop robust active-sampling rules that interpolate between uniform and uncertainty-directed acquisition when model uncertainty may be misspecified, while Xu et al. (2024) combine adaptive label acquisition with predictor-assisted, anytime-valid inference and optimize the labeling policy under a budget. Our problem differs in that every record is already labeled: expert effort is spent not on acquiring missing labels, but rather verifying and potentially correcting that label. Our acquisition objective is not simply where a label is uncertain, but where a possible correction most affects the downstream score, including in nonlinear ways.

8 Conclusion

Our work challenges a common assumption within machine learning communities: that the labels we treat as ground truth are reliable, even though they may be a source of consequential errors. When these labels feed into complex downstream score functions, we show that the structure of these score functions can be leveraged to guide priority-aware auditing, substantially reducing the number of expert audits needed to obtain narrow confidence sequences while retaining finite-sample, distribution-free, and anytime-valid guarantees.

Methodologically, we represent label corrections through a confusion matrix and couple adaptive audit allocation with inverse-probability-weighted inference. Our approach focuses audits on unresolved corrections that matter most to the downstream task without sacrificing validity under adaptive sampling or stopping. The efficiency and validity of this approach is relevant to many high-stakes settings, such as auditing wildlife observations to determine payments in biodiversity credit programs.

Limitations and future work. The gain from priority-aware sampling scales with how unevenly the score-relevant uncertainty is distributed across strata so on very homogeneous data the priority-aware approaches buy little; additionally, we treat expert-verified labels as ground truth, when in reality they are also subject to error. Natural next steps are less conservative confidence sequences that avoid stronger modeling assumptions, and extensions to shifting confusion patterns, heterogeneous audit costs, and multiple auditors.

AI Use Statement

Generative AI tools assisted the sketching and checking of some proofs, literature searches, implementation of experiment and figure generation code, and high-level editing of this manuscript.

Reproducibility Statement

The appendices document the assumptions, proofs, and experimental settings. The supplementary repository provides the implementation, frozen seeds, saved results, figure-generation scripts, and public CIFAR-10N data. The Kenyan biodiversity data are confidential and cannot be redistributed; only anonymized aggregate results and the corresponding experimental code are included.

Ethics Statement

The biodiversity study is a retrospective replay and does not independently validate any credit or payment. CIFAR-10N human annotations are treated as a designated comparison label, not proof that either annotation source is infallible. Prospective use must prespecify the eligible label vocabulary, preserve the recorded randomized design, and account for expert error when verified labels are not definitive.

References

T Mitchell Aide. The biodiversity credit market needs rigorous baseline, monitoring, and validation practices. *npj Biodiversity*, 3(1):30, 2024.

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. 2015.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of Information Theory*, volume 2. wiley New York, 1991.

Donald A Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. chapman and hall/cRc, 1995.

Andrew Gonzalez, Tom August, Sallie Bailey, Kyle Bobiwash, Philipp H Boersch-Supan, Neil D Burgess, Barnabas H Daru, Chris S Elphick, Robert P Freckleton, Winifred F Frick, et al. From data to decisions: Toward a biodiversity monitoring standards framework. *Proceedings of the National Academy of Sciences*, 123(10):e2519347123, 2026.

Sarah R. Haile and David Kronthaler. Potential for bias in prevalence estimates when not accounting for test sensitivity and specificity: A systematic review of COVID-19 seroprevalence studies. *International Journal of Public Health*, 70:1608343, 2025.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

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

Hyun-Chul Kim and Zoubin Ghahramani. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, pp. 619–627. PMLR, 2012.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pp. 5753–5763. PMLR, 2021.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active surrogate estimators: An active learning approach to label-efficient model evaluation. In *Advances in Neural Information Processing Systems*, 2022.

Puheng Li, Tijana Zrnic, and Emmanuel Candes. Robust sampling for active statistical inference. *Advances in Neural Information Processing Systems*, 38:68727–68756, 2025.

Tudor Manole and Aaditya Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 69(7):4641–4658, 2023.

Daksh Mittal, Yuanzhe Ma, Shalmali Joshi, and Hongseok Namkoong. Adaptive labeling for efficient out-of-distribution model evaluation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 70981–71003, 2024.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–606, 1934. doi: 10.1111/j.2397-2335.1934.tb04184.x.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the 35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Charlotte Pelletier, Silvia Valero, Jordi Inglada, Nicolas Champion, Claire Marais Sicre, and Gérard Dedieu. Effect of training class label noise on classification performances for land cover mapping with satellite image time series. *Remote Sensing*, 9(2):173, 2017.

Plan Vivo Foundation. *PV Nature Methodology and Data Protocol*. Plan Vivo Foundation, April 2026. URL https://www.planvivo.org/resource/pv-nature-methodology-data-protocol. Version 1.2.

Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D Tizabi, Nicholas Schreck, A Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein. Labelling instructions matter in biomedical image analysis. *Nature Machine Intelligence*, 5(3):273–283, 2023.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

J Jon Ryu and Gregory W Wornell. Gambling-based confidence sequences for bounded random vectors. *ICML*, 2024.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer Science & Business Media, 2003.

Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active comparison of prediction models. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pp. 39–48, 1974.

Shubhanshu Shekhar, Ziyu Xu, Zachary Lipton, Pierre Liang, and Aaditya Ramdas. Risk-limiting financial audits via weighted sampling without replacement. In *Uncertainty in Artificial Intelligence*, pp. 1932–1941. PMLR, 2023.

Jialin Shi, Kailai Zhang, Chenyi Guo, Youquan Yang, Yali Xu, and Ji Wu. A survey of label-noise deep learning for medical image analysis. *Medical Image Analysis*, 95:103166, 2024.

Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Steven J Sweeney, Kurt C VerCauteren, Nathan P Snow, Joseph M Halseth, Paul A Di Salvo, Jesse S Lewis, Michael D White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019.

Lieven Vandenberghe and Stephen Boyd. *Convex Optimization*, volume 1. Cambridge university press Cambridge, 2004.

Jean Ville. *Etude critique de la notion de collectif*, volume 3. Gauthier-Villars Paris, 1939.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. *Advances in Neural Information Processing Systems*, 33:20204–20214, 2020.

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

Ian Waudby-Smith, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis, and Paul Mineiro. Anytime-valid off-policy inference for contextual bandits. *ACM/IMS Journal of Data Science*, 1(3):1–42, 2024.

World Economic Forum. Biodiversity credits: Demand drivers and guidance on early use. https://www.weforum.org/publications/biodiversity-credits-demand-drivers-and-guidance-on-early-use/, December 2023. Published December 1, 2023; accessed July 29, 2026.

Sven Wunder, Cecilia Fraccaroli, Joseph W Bull, Trishna Dutta, Alison Eyres, Megan C Evans, Bo Jellesmark Thorsen, Julia PG Jones, Martine Maron, Bart Muys, et al. Biodiversity credits: An overview of the current state, future opportunities, and potential pitfalls. *Business Strategy and the Environment*, 34(7):8470–8499, 2025.

Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. Active, anytime-valid risk controlling prediction sets. *Advances in Neural Information Processing Systems*, 37:60110–60132, 2024.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62993–63010, 2024.

A AUDITS FOR BIODIVERSITY CREDIT PROGRAMS

To establish market credibility and prevent greenwashing, voluntary biodiversity credit schemes require standardized monitoring frameworks, robust baseline metrics, and transparent verification processes (Wunder et al., 2025; Gonzalez et al., 2026). A project site is a landowner enrolled in a biodiversity credits program, such as a cattle rancher in rural Kenya. Each project site collects data using sensors placed on their landscape, recording wildlife observations. These data points are labeled with the specific species observed, which determine the biodiversity index for the landscape.

As a representative application, we consider the Plan Vivo "multimetric" (Plan Vivo Foundation, 2026) which includes three species-based pillars[1]: species richness P_1 (number of unique species), species diversity P_2 (relative abundance of species), and taxonomic diversity P_3 (dissimilarity in the taxonomic tree). The multimetric determines the number of biodiversity certificates that the landowner is awarded, which then determines their payment.

Species are categorized into disjoint groups $\mathcal{G} = \{G_1, \dots, G_K\}$, where each group $G_k \subseteq \mathcal{L}$ is a category such as birds, bats, or medium-to-large mammals. The within-group relative abundance of species $s \in G_k$ is

$$p_s^{(G_k)}(C) = \frac{m_s(C)}{\sum_{j \in G_k} m_j(C)}, \qquad s \in G_k. \tag{9}$$

Let $z_s(C) = \mathbb{1}\{m_s(C) > 0\}$ indicate whether species s is present in the corrected labels, $S_k(C) = \sum_{s \in G_k} z_s(C)$ denote the number of species in group k, and $S(C) = \sum_{G_k \in \mathcal{G}} S_k(C)$ denote the total number of present species. The three pillars of the multimetric can then be expressed as:

$$P_1(C) = S(C), \tag{10}$$

$$P_2(C) = \sum_{G_k \in \mathcal{G}} D_1\big(p^{(G_k)}(C)\big), \qquad D_1(p) = \exp\Big(-\sum_s p_s \log p_s\Big), \tag{11}$$

$$P_3(C) = \frac{\sum_{i<j} \omega_{ij} z_i(C) z_j(C)}{\sum_{i<j} z_i(C) z_j(C)} + \sum_{G_k \in \mathcal{G}} \frac{S_k(C)}{S(C)} \Delta\big(p^{(G_k)}(C)\big). \tag{12}$$

Here, the within-group taxonomic dissimilarity is $\Delta(p) = \sum_{i<j} \omega_{ij} p_i p_j / \sum_{i<j} p_i p_j$, where $\omega_{ij} \in [0, \Omega]$ denotes taxonomic distance and Ω is some maximal taxonomic distance. The multimetric then aggregates the three pillar values, e.g., as a weighted sum: $M(C) = \Phi(P_1(C), P_2(C), P_3(C))$. Notably, this scoring function is both nonlinear (P_2 and P_3) and nonsmooth (P_1 and P_3).

[1]The multimetric also includes two habitat pillars. Here, we only consider the species-specific pillars.

B NOTATION

Table 3: Core notation. All submitted and verified labels, and hence the true confusion matrix, are fixed; probability is with respect to the randomized audit design.

Symbol	Meaning
N	Number of submitted records
$y_i \in \mathcal{R} = \{1, \dots, R\}$	Submitted label of record i, known.
$v_i \in \mathcal{L} = \{1, \dots, L\}$	Expert-verified label of record i, unknown and revealed only upon verifying i.
$\mathbf{C}^\star$	True (unknown) confusion matrix, where $C^\star_{rs} = \#\{i : y_i = r, v_i = s\}$.
C, q	A candidate confusion matrix and a candidate complete row-count vector.
N_r, w_r	Known size and dataset weight of submitted-label row r: $N_r = \sum_s C^\star_{rs}$ and $w_r = N_r/N$.
$m_s(C)$, $p_s(C)$	Corrected abundance and corrected proportion of label s induced by C: $m_s(C) = \sum_r C_{rs}$ and $p_s(C) = m_s(C)/N$. $p^\star = p(C^\star)$.
$\theta_r(C) = C_{r\cdot}/N_r$	Candidate transition distribution in row r; $p(C) = \sum_r w_r \theta_r(C)$.
t, u, $u_r(t)$	Global audit time, row-local audit time, and number of randomized audits completed in row r by global time t.
$V_{r,u}$, $A_{rs,u}$, $\mathcal{F}_{r,u}$	Verified label revealed by the u-th randomized audit in row r; $A_{rs,u} = \sum_{\ell \le u} \mathbb{1}\{V_{r,\ell} = s\}$; row-local audit history.
$A_{+s,t}$	Column total of observed verified labels, $\sum_r A_{rs,u_r(t)}$.
R_t, V_t, I_t, $\pi_{r,t}$, $M_{r,t}$	Row chosen at global time t, its revealed label, the reviewed record, the predictable row probability, and the remaining row size $N_r - u_r(t-1)$.
$\mathcal{A}_{t-1}, \mathcal{U}_{t-1}$	Reviewed and remaining record sets before review t (distinct from the count matrix A).
$O_{t,s}$	Reviewed records verified as s by time t
$Z_t(d_t)$, $\widetilde{Z}_t(d_t)$	Row-baseline and metadata-corrected directional observationsSection 3.
$q_s(p)$, $d_s(p)$, c	Smoothed candidate score $(Np_s + 1)/(N + L)$, $-\log q_s(p)$, and slack $\log(1 + L/N)$.
$\alpha_{\text{row}}, \alpha_{\text{coord}}, \alpha_-, \alpha_+$	Declared failure budgets for the matrix event, the composition event, and the two entropy tails.
P_1, P_2, P_3	The three species-based pillars: richness, diversity, and taxonomic dissimilarity.

C ALGORITHM PSEUDOCODE

Reading algorithm 1. All predictions, priorities, sampling probabilities, and betting stakes are computed before the next verified label is observed; this predictability is what permits adaptive allocation without sacrificing validity. Verified strata are removed and the sampling distribution is renormalized over the remaining records. Prediction quality affects efficiency only: an inaccurate predictor may slow contraction but does not invalidate the confidence sequence.

Reading algorithm 2. The two entropy endpoints require different certificates. The upper endpoint uses a predictable linear cross-entropy direction, whereas the lower endpoint must rule out every composition still compatible with the audits through a conservative convex optimization. Both endpoints use the same audit stream, with tail budgets satisfying $\alpha_+ + \alpha_- \le \alpha$; the reported interval is valid only when the numerical lower-end optimization returns a certified bound rather than an unconstrained approximate solution.

Algorithm 1 Priority-Aware Sampling (PAS) for a linear score

Require: Strata g, contribution matrix $\mathbf{D}$, predictor PREDICT(), miscoverage α, SRS weight ε, target half-width τ

1: **while** confidence-sequence half-width $> \tau$ **do**

2: **Predict:** For each nonempty stratum g, estimate $q_{g,t}$ and

$$\widehat{\mu}_{g,t} = \sum_s q_{g,t}(s) D_{r(g),s}, \qquad \widehat{\sigma}^2_{g,t} = \sum_s q_{g,t}(s) \big(D_{r(g),s} - \widehat{\mu}_{g,t}\big)^2.$$

3: **Estimate score:**

$$\widehat{\theta}_{t-1} = \frac{1}{N} \left[\sum_{i \in \mathcal{A}_{t-1}} D_{y_i, v_i} + \sum_g M_{g,t} \widehat{\mu}_{g,t} \right].$$

4: **Update sampling probabilities:**

$$\pi_{g,t} = \varepsilon \frac{M_{g,t}}{\sum_k M_{k,t}} + (1-\varepsilon) \frac{M_{g,t} \widehat{\sigma}_{g,t}}{\sum_k M_{k,t} \widehat{\sigma}_{k,t}}.$$

▷ Mix residual-record SRS with variance-based priority.

5: **Set stakes:** Compute the predictable range of the next Z_t over all possible strata and verified labels; choose capped stakes $\lambda_t^{\pm}(m)$ for each candidate m.

6: **Audit:** Draw $G_t \sim \pi_t$ and uniformly audit one unaudited record in G_t, revealing V_t.

7: **Observe:**

$$Z_t = \widehat{\theta}_{t-1} + \frac{M_{G_t,t}}{N \pi_{G_t,t}} \left(D_{r(G_t),V_t} - \widehat{\mu}_{G_t,t} \right).$$

8: **Update e-processes:**

$$E_t^+(m) = E_{t-1}^+(m) \big[1 + \lambda_t^+(m)(Z_t - m)\big],$$

$$E_t^-(m) = E_{t-1}^-(m) \big[1 - \lambda_t^-(m)(Z_t - m)\big].$$

9: **Report:**

$$\mathcal{I}_t = \left\{ m : E_s^{\pm}(m) < 2/\alpha \; \forall s \le t \right\},$$

and update the audited set and remaining stratum sizes.

10: **end while**

11: **return** $\mathcal{I}_t$

Algorithm 2 Distribution-free audit of entropy

Require: Strata g, predictor PREDICT(), tail budgets $\alpha_{\pm}$, smoothing η, target log-width $\tau = \log(21/19)$

1: **while** $U_{t-1} - L_{t-1} > \tau$ **do**

2: **Predict composition:** Estimate $q_{g,t}$ for each nonempty stratum and form

$$\widehat{p}_{t-1} = \frac{1}{N}\left[n_{t-1} + \sum_g M_{g,t} q_{g,t}\right].$$

3: **Set entropy direction:**

$$r_t = (1-\eta)\widehat{p}_{t-1} + \eta/L, \qquad \psi_t = -\log r_t.$$

4: **Estimate stratum influence:**

$$b_{g,t} = q_{g,t}^{\top}\psi_t, \qquad \widehat{\sigma}^2_{g,t} = \sum_s q_{g,t}(s)(\psi_{t,s} - b_{g,t})^2.$$

5: **Update sampling probabilities:**

$$\pi_t = 0.4\, p_t^{\mathrm{prop}} + 0.3\, p_t^{\mathrm{floor}} + 0.3\, p_t^{\mathrm{pri}},$$

where

$$p_{g,t}^{\mathrm{prop}} \propto M_{g,t}, \qquad p_{g,t}^{\mathrm{floor}} \propto M_{g,t}(1 + b_{g,t}), \qquad p_{g,t}^{\mathrm{pri}} \propto M_{g,t}\widehat{\sigma}_{g,t}.$$

6: **Set stakes:** Choose predictable upper/lower stakes from the range of ψ_t, the possible Horvitz–Thompson weights, and current endpoints L_{t-1}, U_{t-1}.

7: **Audit:** Draw $G_t \sim \pi_t$ and uniformly audit one unaudited record in G_t, revealing V_t.

8: **Upper endpoint:** Form the prediction-corrected directional observation Z_t^U using contributions ψ_t, update

$$E_t^U(h) = E_{t-1}^U(h)\big[1 - \lambda_t^U(h)(Z_t^U - h)\big],$$

and set U_t to the smallest h ruled out by the upper e-process.

9: **Lower endpoint:** Form (X_t, K_t) and update

$$F_t(p;h) = F_{t-1}(p;h) + \log\big[1 + \lambda_t^L(h)\left\{X_t^{\top}\log(1/\widetilde{p}) + K_t - h - c\right\}\big],$$

with

$$\widetilde{p} = \frac{Np + \mathbf{1}}{N+L}, \qquad c = \log(1 + L/N).$$

10: **Profile over feasible compositions:** For each h, compute a certified lower bound on

$$\min_{p \in \mathcal{P}_t} F_t(p;h)$$

by projected gradient plus a supporting-plane bound; use bisection in h to obtain L_t.

11: **Report:**

$$P_2 \in [e^{L_t}, e^{U_t}], \qquad \widehat{P}_{2,t} = \frac{2e^{L_t+U_t}}{e^{L_t}+e^{U_t}},$$

then update counts and remaining records.

12: **end while**

13: **return** $[e^{L_t}, e^{U_t}]$ and $\widehat{P}_{2,t}$

D Experiment Settings and Additional Results

All error-rate experiments use nominal confidence 95.5% ($\alpha = 0.045$). PAS and SRS use the same prediction-corrected betting confidence sequence and differ only in their predictable sampling probabilities. Records are sampled uniformly within the selected stratum. Paired policies use identical randomized within-stratum record orders.

Table 4: Frozen settings for the label-error-rate experiments. The prior column gives the prior probability that a submitted label is correct. The SRS weight is ε in Equation (7).

Setting	Kenya Dataset	CIFAR-10N
Population size	33,403	50,000
Audit strata	46	40
Audit paths per policy	20	20
Maximum audits	5,000	5,000
Stored interval spacing	25	25
Target half-width	0.02	0.02
SRS weight ε	0.30	0.10
Submitted-label prior	0.80	development-calibrated
Prior/shrinkage strength	2 and 1	50
Stake effective horizon	800	1,200
Stake cap fraction	0.90	0.90
Candidate-grid spacing	5×10^{-4}	5×10^{-4}

D.1 Kenya Biodiversity Setting

D.1.1 Linear Score: Mislabel Rate

Submitted labels define the 46 audit strata. The predictable error model starts from correctness probability 0.8. Observed row error rates are shrunk toward the corresponding error probability 0.2 with strength 2, followed by a second shrinkage stage of strength 1. PAS mixes 30% remaining-record SRS with 70% predicted Neyman allocation. We use 20 paired paths with seeds $0, \ldots, 19$.

D.1.2 Nonlinear Score: Pillar 2 Diversity

As in the main text, let

$$m_s(C) = \sum_r C_{rs}, \qquad p_s(C) = \frac{m_s(C)}{N}$$

denote the corrected abundance and population proportion of species s induced by a candidate confusion matrix C. Pillar 2 is the Hill–Shannon diversity of this corrected composition:

$$H(p(C)) = -\sum_{s \in \mathcal{L}} p_s(C) \log p_s(C), \qquad P_2(C) = \exp\{H(p(C))\}\,.$$

Thus, P_2 is the effective number of species implied by the corrected labels: it increases when the composition contains more species or distributes mass more evenly among them. Label corrections affect P_2 through the column totals of C, and its nonlinearity makes uncertainty in rare or poorly audited species especially consequential.

The distribution-free P_2 experiment uses 300 nonempty strata obtained by crossing submitted label with camera station. The prediction hierarchy shrinks station-level distributions toward their submitted-label row with strength 1, and row distributions toward a pooled composition with strength 2. The row prior places probability 0.8 on the submitted label. The predictable reference composition used in the entropy direction is

$$r_t = 0.99\,\widehat{p}_t + \frac{0.01}{L}\mathbf{1}.$$

PAS uses the frozen mixture

$$\pi_t^{\mathrm{PAS}} = 0.4\,\pi_t^{\mathrm{prop}} + 0.3\,\pi_t^{\mathrm{range}} + 0.3\,\pi_t^{\mathrm{Neyman}}.$$

The proportional component samples according to remaining stratum size, the range component controls the possible range of the importance-weighted observation, and the Neyman component uses predicted within-stratum variation in the current entropy direction.

Table 5: Frozen settings for the Kenyan biodiversity data distribution-free diversity experiment.

Setting	Value
Audit strata	300
Paired paths	20
Seeds	$95001, \ldots, 95020$
Maximum audits	3,000
Endpoint update spacing	25
Reported checkpoint spacing	100
Lower-tail error budget	0.0225
Upper-tail error budget	0.0225
Proportional/range/Neyman weights	$0.40/0.30/0.30$
Reference smoothing	0.01
Row/stratum shrinkage	$2/1$
Stake cap fraction	0.90
Entropy resolution	3×10^{-4}
Profile iteration limit	60
Target entropy width	$\log(21/19)$

The entropy interval is transformed outward to the P_2 scale. The stopping target $U_t^H - L_t^H \leq \log(21/19)$ corresponds to at most 5% relative uncertainty around the harmonic center of the transformed endpoints. SRS and PAS use the same lower convex-profile certificate, upper affine certificate, prediction hierarchy, and stake rules.

D.1.3 Bayesian Settings

We evaluate the model-based procedure of section 6 on the Kenyan biodiversity dataset. PAS and SRS use the same hierarchical Dirichlet–multinomial model, posterior computation, credible interval, and stopping rule; they differ only in audit allocation. Records are sampled uniformly within the selected stratum, and paired PAS and SRS paths use the same randomized within-stratum record orders.

The model hyperparameters are fixed at $\beta = 0.5$, $\kappa_r = 2$, and $\kappa_c = 0.5$, with add-one smoothing of the pooled verified-label composition. Posterior intervals are evaluated every 25 audits beginning at audit 100. Each evaluation uses 400 posterior draws and reports the 2.25th and 97.75th percentiles of P_2.

PAS uses the frozen allocation

$$\pi_t^{\mathrm{PAS}} = 0.2\,\pi_t^{\mathrm{prop}} + 0.8\,\pi_t^{\mathrm{Neyman}}, \qquad \pi_{g,t}^{\mathrm{Neyman}} \propto M_{g,t}\hat{\sigma}_{g,t},$$

where $\hat{\sigma}_{g,t}$ is the predicted within-stratum standard deviation of the entropy influence direction. Its variance includes the light-audit adjustment $1 + 1/(n_{g,t} + \kappa_c)$. SRS samples uniformly from all unaudited records. Neither policy uses a pilot phase.

D.2 CIFAR-10N Setting

D.2.1 Linear Score: Mislabel Rate

The submitted label is `random_label1`, and `clean_label` is the designated audit reference. The two additional human annotations, `random_label2` and `random_label3`, are used as pre-audit covariates. Each record belongs to one of 40 strata formed by crossing its submitted class with four auxiliary agreement patterns: both auxiliary labels support the submission, exactly one supports it, both agree on a different label, or all three human annotations differ.

Stratum error priors were estimated on a development view using `random_label2` as the submitted label and the other two human rounds as auxiliary annotations. Six candidate policies were evaluated on five development paths through 1,500 audits. The frozen selection rule minimized median terminal

Table 6: Frozen settings for the Bayesian diversity experiment.

Setting	Value
Audit strata	300
Audit paths per policy	100
Seeds	$8000, \ldots, 8099$
Maximum audits	1,000
Evaluation schedule	$100, 125, \ldots, 1000$
Posterior draws per evaluation	400
Credible-interval percentiles	2.25 / 97.75
Prior weight β	0.5
Concentrations κ_r/κ_c	$2/0.5$
Pooled-composition smoothing	Add one
PAS proportional/Neyman weights	$0.20/0.80$
Pilot audits	0
Stopping threshold	$U_t^H - L_t^H \leq \log(21/19)$

half-width, breaking ties by mean half-width and then by configuration identifier. The selected policy uses a prior strength of 50, a stake effective horizon of 1,200, and a 10% SRS component. Confirmation uses 20 fresh paired paths with seeds $2026101200, \ldots, 2026101219$.

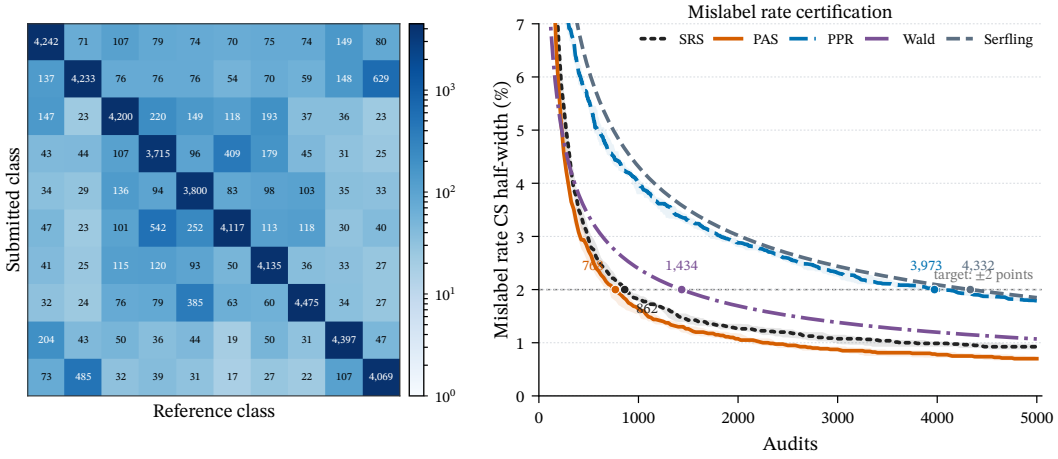

Figure 4: Audits on CIFAR-10N. **(Left)** Submitted-to-reference confusion matrix. Diagonal entries indicate agreement with the reference labels, while off-diagonal entries indicate label disagreements. Cell annotations give record counts, and color displays them on a log scale. **(Right)** Half-width of the 95.5% mislabel-rate confidence sequence as the number of audits increases. Methods and plotting conventions match fig. 2; dots mark the audits required to reach a half-width of two percentage points. The two settings exhibit similar contraction patterns and relative method performance. Numerical stopping times for both are reported in table 1.

D.2.2 Nonlinear Score: Worst-Three Scores

For a confusion matrix C, let $N_r = \sum_{s=1}^{10} C_{rs}$ denote the number of records submitted as class r. The classwise precision is $p_r(C) = C_{rr}/N_r$. We define worst-three precision as the mean precision of the three least precise submitted classes:

$$g_{\text{prec}}(C) = \min_{\substack{S \subseteq [10] \\ |S|=3}} \frac{1}{3} \sum_{r \in S} \frac{C_{rr}}{N_r}.$$

To define taxonomic risk, we map each CIFAR-10 class to a fixed noun synset in Princeton WordNet 3.0. For each pair of classes r, s, we compute the unweighted shortest-path distance between their

synsets in the undirected WordNet hypernym graph and normalize it by the largest distance among the ten classes. This gives a fixed semantic-distance matrix $D = (d_{rs})$, with $d_{rr} = 0$ and $d_{rs} \in [0, 1]$.

The classwise taxonomic risk is the average semantic severity of corrections from submitted class r,

$$\tau_r(C) = \frac{1}{N_r} \sum_{s=1}^{10} C_{rs} d_{rs}.$$

We define worst-three taxonomic risk as the mean risk of the three highest-risk submitted classes:

$$g_{\text{tax}}(C) = \max_{\substack{S \subseteq [10] \\ |S|=3}} \frac{1}{3} \sum_{r \in S} \tau_r(C) = \max_{\substack{S \subseteq [10] \\ |S|=3}} \frac{1}{3} \sum_{r \in S} \frac{1}{N_r} \sum_{s=1}^{10} C_{rs} d_{rs}.$$

Both scores are therefore the minimum or maximum of the $\binom{10}{3} = 120$ fixed three-row averages. The candidate-score grid has spacing 0.001, the total miscoverage level is $\alpha = 0.045$, the stake effective horizon is 800, and the stake cap fraction is 0.9. Predictions use a submitted-label prior of 0.8, row shrinkage strength 2, and secondary shrinkage strength 1.

PAS uses

$$\pi_t^{\text{PAS}} = 0.70\, \pi_t^{\text{prop}} + 0.15\, \pi_t^{\text{range}} + 0.15\, \pi_t^{\text{Neyman}}.$$

Row relevance gives equal weight to plug-in worst-three membership and to the three-row subsets currently limiting the profiled endpoint. The latter weights are smoothed with softmax temperature 1, and ties in estimated tail membership are divided fractionally.

D.3 Synthetic Data: A Priority Mechanism Study

The synthetic experiment isolates the acquisition advantage produced by heterogeneous score relevance. Its confirmation population contains 20 equal-sized submitted-label rows with 250 records per row, for $N = 5{,}000$. Four rows have equal numbers of records with score contributions 0 and 1. The other 16 rows each have two possible verified destinations, both with score contribution 0.25. Their verified labels remain uncertain, but that uncertainty does not affect the chosen score.

The target is the population-wide linear risk

$$g(C) = \frac{1}{N} \sum_{r,s} A_{rs} C_{rs},$$

whose value in the confirmation population is 0.30. PAS uses the remaining-size-aware Neyman rule

$$\pi_{r,t}^{\text{PAS}} \propto M_{r,t} \widehat{\sigma}_{r,t}$$

without an SRS floor. Constant-contribution rows have zero predicted score variance and therefore receive no PAS audits. SRS remains uniform over all unaudited records. Both policies use the same prediction-assisted linear betting confidence sequence.

The scaling screen fixes four score-variable rows and varies the total number of rows over

$$R \in \{8, 12, 20, 32, 48\}.$$

Each configuration uses five development seeds and an 800-audit budget. This screen tests the predicted increase in PAS advantage as the number of score-invariant nuisance rows grows. Confirmation fixes $R = 20$ and uses the 100 fresh paths reported above.

D.4 Random-Sampling Comparators

The default prior–posterior-ratio confidence sequence is evaluated on the same 20 SRS audit orders used by our betting confidence sequence. It uses a uniform $\text{BetaBinomial}(N, 1, 1)$ working prior on the population error count and reports the running intersection of its candidate sets.

The Wald reference uses the true population error variance, no finite-population correction, and nominal confidence 95.5%. It is an oracle-variance, fixed-sample asymptotic reference. The Serfling reference is the corresponding finite-population, fixed-sample bound. Wald and Serfling therefore provide benchmark sample sizes rather than anytime-valid stopping rules.

Table 7: Frozen synthetic-experiment settings.

Setting	Value
Confirmation rows	20
Score-variable rows	4
Records per row	250
Population size	5,000
Confirmation paths	100
Confirmation seeds	2026101000, . . . , 2026101099
Maximum audits	800
Checkpoint spacing	25
Confidence level	95.5%
Prior strength	2
Stake effective horizon	800
Stake cap fraction	0.90
Candidate-grid spacing	5×10^{-4}
SRS component in PAS	0

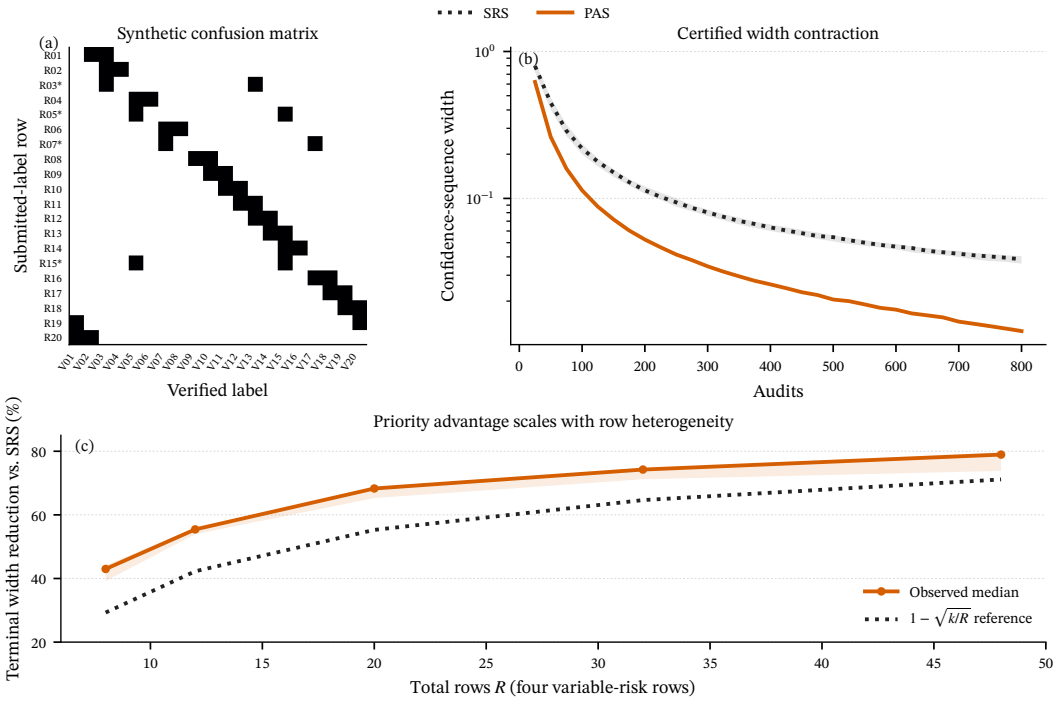

Figure 5: Controlled synthetic experiment isolating the benefit of priority-aware sampling. **(a)** Submitted-to-verified confusion matrix; asterisks identify the four rows whose possible verified labels have different score contributions. **(b)** Median confidence-sequence width under PAS and SRS, with interquartile ranges. **(c)** Terminal PAS width reduction as the total number of rows increases while the number of score-variable rows remains fixed at four. The dotted curve shows the $1 - \sqrt{k/R}$ reference with $k = 4$.

E Computational Environment

All experiments were executed on a local machine with 15 CPU cores and 24 GB of unified memory. Computation was CPU-only; no GPU or external computing cluster was used. Independent score-policy-seed tasks were distributed across four local worker processes.

The final runs used macOS 26.5 and Python 3.14.7. Principal numerical and reporting dependencies were NumPy 2.5.2, SciPy 1.18.1, pandas 3.0.5, and Matplotlib 3.11.1. Each worker wrote an

independent result shard, and aggregation began only after all expected shards for that experiment had completed.

F LINEAR SCORES

Lemma 4 (Ville's inequality Ville (1939)). *Let $(W_t)_{t\geq 0}$ be a nonnegative supermartingale with respect to $(\mathcal{F}_t)$ with $W_0 = 1$. Then $\Pr\{\exists t : W_t \geq c\} \leq 1/c$ for every $c > 0$.*

Predictions. Any predictable $q_{g,t}$ keeps Lemma 1 valid; good predictions only reduce the variance (5). We use a two-level shrinkage estimator, built only from the audits before audit t. Let $A_{r,t}(s)$ and $A_{g,t}(s)$ be the numbers of audited records in row r and in stratum g that were verified as s, and $n_{r,t} = \sum_s A_{r,t}(s)$, $n_{g,t} = \sum_s A_{g,t}(s)$ the numbers audited. Then

$$\bar{g}_t(s) = \frac{1/L + \sum_r A_{r,t}(s)}{1 + (t-1)} \qquad \text{(pooled verified labels)}$$

$$q_{r,t} = \frac{A_{r,t} + \kappa_r\{\beta e_r + (1-\beta)\bar{g}_t\}}{n_{r,t} + \kappa_r} \qquad \text{(submitted label } r\text{)} \tag{13}$$

$$q_{g,t} = \frac{A_{g,t} + \kappa_c\, q_{r(g),t}}{n_{g,t} + \kappa_c} \qquad \text{(stratum } g \text{ within row } r(g)\text{)}, \tag{14}$$

where e_r is the unit vector at r. Before any audit in row r, the prediction puts weight β on "the submitted label is correct" and spreads the rest over the labels verified so far. As audits accumulate, each row moves toward its own verified-label frequencies, and each stratum toward its own frequencies within its row. Each level is the posterior mean of a Dirichlet–multinomial model whose prior is centred on the level above. This is the same structure as the Bayesian model of section 6, but here it is used only for prediction, never for inference. On the Kenya data we use $\beta = 0.8$, $\kappa_r = 2$ and $\kappa_c = 1$, fixed before the audit and not tuned.

F.1 PROOFS FOR SECTION 3

Proof of lemma 1. Unbiasedness. Given $\mathcal{F}_{t-1}$ and $G_t = g$, I_t is uniform on $\mathcal{U}_{g,t-1}$, so $\mathbb{E}[D_{y_{I_t}V_t} \mid \mathcal{F}_{t-1}, G_t = g] = \mu_{g,t}$. Averaging over G_t,

$$\begin{aligned}\mathbb{E}\Big[\frac{M_{G_t,t}}{N\pi_{G_t,t}}\big(D_{y_{I_t}V_t} - q_{G_t,t}^\top D_{r(G_t)\cdot}\big)\;\Big|\;\mathcal{F}_{t-1}\Big] &= \sum_{g:M_{g,t}>0} \pi_{g,t}\frac{M_{g,t}}{N\pi_{g,t}}\big(\mu_{g,t} - q_{g,t}^\top D_{r(g)\cdot}\big)\\ &= \frac{1}{N}\sum_g M_{g,t}\big(\mu_{g,t} - q_{g,t}^\top D_{r(g)\cdot}\big).\end{aligned}$$

The condition $\pi_{g,t} > 0$ whenever $M_{g,t} > 0$ lets the probabilities cancel. Adding $\hat{\theta}_{t-1}$, the predicted terms cancel and (2) gives $\theta^\star$. *Variance.* Only the correction is random. Its conditional second moment is $\sum_g \pi_{g,t}(M_{g,t}/N\pi_{g,t})^2\mathbb{E}[(D_{y_{I_t}V_t} - q_{g,t}^\top D_{r(g)\cdot})^2 \mid G_t = g] = N^{-2}\sum_g (M_{g,t}^2/\pi_{g,t})(\sigma_{g,t}^2 + b_{g,t}^2)$. Subtracting the square of its mean, $N^{-1}\sum_g M_{g,t}b_{g,t}$, gives (5). □

Proof of proposition 3. With exact predictions $b_{g,t} = 0$ and (5) is $N^{-2}\sum_g \beta_g^2/\pi_g$ with $\beta_g = M_{g,t}\sigma_{g,t}$. This is convex on $\{\pi \geq \ell, \sum_g \pi_g = 1\}$, and the KKT conditions give $\pi_g = \max\{\ell_g, \beta_g/\nu\}$. Without floors, Cauchy–Schwarz gives $\sum_g \beta_g^2/\pi_g \geq (\sum_g \beta_g)^2$ with equality at $\pi \propto \beta$. For refinement, split g into $g_1, \ldots, g_J$ with shares ρ_j; by the law of total variance and Jensen, $\sigma_g^2 \geq \sum_j \rho_j \sigma_{g_j}^2 \geq (\sum_j \rho_j \sigma_{g_j})^2$, so $M_g\sigma_g \geq \sum_j M_{g_j}\sigma_{g_j}$. □

The range of one observation. The stake caps need predictable bounds valid for every stratum that could be drawn and every unobserved label:

$$\underline{Z}_t = \hat{\theta}_{t-1} + \min_{g:M_{g,t}>0} \frac{M_{g,t}}{N\pi_{g,t}}\Big(\min_v D_{r(g)v} - q_{g,t}^\top D_{r(g)\cdot}\Big),$$

$$\overline{Z}_t = \hat{\theta}_{t-1} + \max_{g:M_{g,t}>0} \frac{M_{g,t}}{N\pi_{g,t}} \Big(\max_v D_{r(g)v} - q_{g,t}^\top D_{r(g)\cdot} \Big).$$

For the error rate, $\min_v D_{rv} = 0$ and $\max_v D_{rv} = 1$ in every row.

Proof of theorem 2. Consider E_t^+ at $m = \theta^\star$. If $\theta^\star > \underline{Z}_t$ the stake is at most $0.9/(\theta^\star - \underline{Z}_t)$, so $1 + \lambda_t^+(Z_t - \theta^\star) \geq 1 - 0.9 = 0.1$; if $\theta^\star \leq \underline{Z}_t$ the factor is at least one. By lemma 1 its conditional mean is one. So $E_t^+(\theta^\star)$ is a nonnegative martingale with $E_0^+ = 1$, and lemma 4 gives $\Pr\{\exists t : E_t^+(\theta^\star) \geq 2/\alpha\} \leq \alpha/2$. The same holds for $E_t^-(\theta^\star)$ with the cap $0.9/(\overline{Z}_t - \theta^\star)$. The sequence misses $\theta^\star$ only if one of the two reaches $2/\alpha$. □

Evaluating the sequence on a grid. In the experiment, candidates are evaluated on the grid $m_j = j/2000$, $j = 0, \ldots, 2000$. On each interval $[m_j, m_{j+1}]$ each side's stake is held constant at the more constrained endpoint: λ^+ capped at $0.9/(m_{j+1} - \underline{Z}_t)$ and λ^- at $0.9/(\overline{Z}_t - m_j)$. Then $1+\lambda^+(Z_t - m)$ decreases and $1-\lambda^-(Z_t - m)$ increases in m on the interval, so if $\theta^\star \in [m_j, m_{j+1}]$ is not rejected, neither are m_{j+1} by the upward test nor m_j by the downward test. Retaining every grid interval whose right end survives the upward test and whose left end survives the downward test therefore preserves coverage exactly, and differs from the hull of surviving grid points by at most one grid step (only 0.05 percentage points).

Choosing the stakes. Within the caps, the stake decides how fast evidence accumulates. Suppose the observations have mean $m + \delta$ and variance v. Expanding $\log(1 + x) \approx x - x^2/2$, each review adds about $\lambda\delta - \lambda^2 v/2$ to $\log E_t^+(m)$, so after n reviews the candidate m is ruled out once $n(\lambda\delta - \lambda^2 v/2) \geq \log(2/\alpha)$, that is, once $\delta \geq \log(2/\alpha)/(n\lambda) + \lambda v/2$. The stake that makes this smallest is

$$\lambda = \sqrt{\frac{2\log(2/\alpha)}{n\,v}}, \qquad \text{giving} \qquad \delta = \sqrt{\frac{2v\log(2/\alpha)}{n}}.$$

A large stake is safe when the observations are quiet and dangerous when they are noisy. Neither v nor the number of reviews is known in advance, so we use the predictable plug-in rule of Waudby-Smith & Ramdas (2024) in its fixed-horizon form: we replace v by a running estimate computed from past reviews only, and n by a planned horizon $n^\star$,

$$\lambda_t = \sqrt{\frac{2\log(2/\alpha)}{n^\star \hat{v}_t}}, \qquad \lambda_t^+(m) = \min\Big\{\lambda_t, \frac{0.9}{m - \underline{Z}_t}\Big\}, \qquad \lambda_t^-(m) = \min\Big\{\lambda_t, \frac{0.9}{\overline{Z}_t - m}\Big\}, \tag{15}$$

where a cap is dropped when its denominator is nonpositive and

$$\hat{v}_t = \frac{1}{t}\Big\{\tfrac{1}{4} + \sum_{i<t}\big(Z_i - \hat{\theta}_{i-1}^Z\big)^2\Big\}, \qquad \hat{\theta}_i^Z = \frac{1}{i}\sum_{j\leq i} Z_j, \qquad \hat{\theta}_0^Z = \hat{\theta}_0.$$

The term $\frac{1}{4}$, the largest variance of a $[0,1]$ variable, keeps early stakes small before any variance has been observed. The horizon $n^\star$ is the number of reviews at which the stakes are sharpest. The half-width of the interval is insensitive to $n^\star$, and we use $n^\star = 800$ throughout without tuning.

This is what separates the e-process from Serfling's bound. Serfling's width is set by the largest possible variance, $1/4$; here the width $\{2\hat{v}_t \log(2/\alpha)/n\}^{1/2}$ is set by the variance actually observed, which is 0.147 for the raw error indicator and far smaller once the predictions of (4) have learned the rows.

G Entropy and linear envelopes: construction and proofs

G.1 Target and guarantee

We write $p^\star = p(\mathbf{C}^\star)$ for the corrected composition, $H^\star = H(p^\star)$ for its entropy, and Δ^{L-1} for the set of compositions on the vocabulary $\mathcal{L}$, with $0\log 0 = 0$. The audit reports a lower end L_t^H and an upper end U_t^H after every audit. We split the total level as $\alpha = \alpha_- + \alpha_+$, and require

$$\Pr\big\{L_t^H \leq H^\star \leq U_t^H \text{ for all } t \geq 0\big\} \;\geq\; 1 - \alpha_- - \alpha_+ \qquad \text{for every labeling.} \tag{16}$$

Here α_- bounds the probability that the lower end ever exceeds $H^\star$, and α_+ bounds the probability that the upper end ever falls below it. The experiments use $\alpha_- = \alpha_+ = 0.0225$.

The diversity interval is $P_2 \in [e^{L_t^H}, e^{U_t^H}]$. We stop once the report is within 5% of every value in the interval, which by proposition 5 means $U_t^H - L_t^H \le \log(21/19)$.

Proposition 5 (Reporting a point value)**.** *For $0 < \underline{P} \le \overline{P}$, the report minimizing $\max_{P\in[\underline{P},\overline{P}]} |x - P|/P$ is $x^\star = 2\underline{P}\,\overline{P}/(\underline{P} + \overline{P})$. Its worst-case relative error is $(\overline{P} - \underline{P})/(\overline{P} + \underline{P})$, which equals $\tanh\{(U_t^H - L_t^H)/2\}$ for $\underline{P} = e^{L_t^H}$ and $\overline{P} = e^{U_t^H}$. So 5% relative accuracy holds if and only if $U_t^H - L_t^H \le 2\,\mathrm{artanh}(0.05) = \log(21/19)$.*

Proof. For $x \in [\underline{P}, \overline{P}]$ the worst case is at an endpoint, giving $\max\{x/\underline{P} - 1,\ 1 - x/\overline{P}\}$. A report outside the interval is worse at the nearer endpoint. The first term increases in x and the second decreases; they meet at $x^\star$ with value $(\overline{P} - \underline{P})/(\overline{P} + \underline{P})$. Writing $W = U_t^H - L_t^H$, this equals $(e^W - 1)/(e^W + 1) = \tanh(W/2)$, and $\tanh(W/2) \le 0.05$ if and only if $W \le \log(1.05/0.95)$. □

G.2 The reference composition and the observation from each audit

Gibbs' inequality. For a reference r with every $r_s > 0$, $\mathrm{CE}(p, r) - H(p) = \mathrm{KL}(p\|r)$. By Jensen's inequality,

$$-\mathrm{KL}(p\|r) = \sum_{s:p_s>0} p_s \log(r_s/p_s) \le \log \sum_{s:p_s>0} r_s \le 0,$$

with equality if and only if $r = p$ (Cover et al., 1991, Theorem 2.6.3). Since $\partial_s H(r) = -\log r_s - 1$, formula (8) gives $d_s(r) = \log(1/r_s)$.

The plug-in composition and the reference. Let $O_{t,s}$ be the number of the first t audited records verified as s, and let $q_{g,t}(s)$ be the predicted share of label s among the unaudited records of stratum g (the predictions of section 3). Before audit t,

$$\hat{p}_{t-1,s} = \frac{1}{N}\Big(O_{t-1,s} + \sum_g M_{g,t}\, q_{g,t}(s)\Big), \qquad r_t = (1-\eta)\hat{p}_{t-1} + \frac{\eta}{L}, \qquad \psi_{t,s} = \log\frac{1}{r_{t,s}}, \tag{17}$$

with $\eta = 0.01$. Four properties follow:

- $\sum_s \hat{p}_{t-1,s} = 1$, because $\sum_s O_{t-1,s} = t - 1$ and $\sum_g M_{g,t} = N - t + 1$;
- $\hat{p}_{t-1}$ equals $p^\star$ once every record has been audited;
- mixing in the uniform distribution gives $\eta/L \le r_{t,s} \le 1$, so every contribution $\psi_{t,s}$ lies in $[0, \log(L/\eta)]$, even for labels not yet seen;
- any predictable reference gives a valid upper end, and the choice of r_t affects only the width, through $\mathrm{KL}(p^\star\|r_t)$, which is second order in the error of $\hat{p}_{t-1}$.

The predicted contribution of stratum g is $\hat{\mu}_{g,t} = \sum_s q_{g,t}(s)\,\psi_{t,s}$.

The observations. Audit t gives the upper-end observation, which is (4) with every row of the contribution matrix equal to ψ_t:

$$Z_t^U = \underbrace{\sum_s \hat{p}_{t-1,s}\,\psi_{t,s}}_{\text{plug-in cross-entropy}} + \frac{M_{G_t,t}}{N\pi_{G_t,t}}\big(\psi_{t,V_t} - \hat{\mu}_{G_t,t}\big). \tag{18}$$

For a candidate composition p, let $\tilde{p} = (Np + 1)/(N + L)$. The same audit gives the lower-end observation

$$Z_t(p) = \frac{1}{N}\Big(\sum_{i\in A_{t-1}} \log\frac{1}{\tilde{p}_{v_i}} + \sum_g M_{g,t}\hat{\mu}_{g,t}\Big) + \frac{M_{G_t,t}}{N\pi_{G_t,t}}\Big(\log\frac{1}{\tilde{p}_{V_t}} - \hat{\mu}_{G_t,t}\Big). \tag{19}$$

This is (18) with $\log(1/\tilde{p}_s)$ in place of $\psi_{t,s}$. The prediction term $\hat{\mu}_{g,t}$ is kept the same for every candidate, which is what makes the computation in section G.5 convex. The smoothing keeps every contribution finite, at most $\log(N + L)$.

Lemma 6 (One audit, every cross-entropy). *If $\pi_{g,t} > 0$ whenever $M_{g,t} > 0$, then $\mathbb{E}[Z_t^U \mid \mathcal{F}_{t-1}] = \mathrm{CE}(p^\star, r_t) = H^\star + \mathrm{KL}(p^\star \| r_t) \geq H^\star$, and $\mathbb{E}[Z_t(p) \mid \mathcal{F}_{t-1}] = \mathrm{CE}(p^\star, \tilde{p})$ for every composition p.*

Proof. Let $\vartheta \in (0,1]^L$ be predictable, let the $\hat{\mu}_{g,t}$ be any predictable numbers, and set

$$Y_t = \frac{1}{N}\Big(\sum_{i \in A_{t-1}} \log \frac{1}{\vartheta_{v_i}} + \sum_g M_{g,t} \hat{\mu}_{g,t} \Big) + \frac{M_{G_t,t}}{N \pi_{G_t,t}} \Big(\log \frac{1}{\vartheta_{V_t}} - \hat{\mu}_{G_t,t} \Big).$$

Given $\mathcal{F}_{t-1}$ and $G_t = g$, the audited record is uniform over the $M_{g,t}$ unaudited records of g. Averaging over G_t cancels $\pi_{g,t}$ exactly as in the proof of lemma 1, and the $\hat{\mu}$ terms cancel. So $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = N^{-1} \sum_{i=1}^N \log(1/\vartheta_{v_i}) = \mathrm{CE}(p^\star, \vartheta)$. The argument uses only that $\hat{\mu}_{g,t}$ is predictable.

With $\vartheta = r_t$ we get $Y_t = Z_t^U$, because (17) gives

$$\frac{1}{N}\Big(\sum_{i \in A_{t-1}} \psi_{t,v_i} + \sum_g M_{g,t} \hat{\mu}_{g,t} \Big) = \sum_s \hat{p}_{t-1,s} \psi_{t,s}.$$

With $\vartheta = \tilde{p}$ we get $Y_t = Z_t(p)$. □

G.3 The upper end

For each candidate value h, the upper-end e-process is

$$E_t^U(h) = \prod_{i \leq t} \{1 - \lambda_i^U (Z_i^U - h)\},$$

the downward e-process of section 3 applied to (18). The stake $\lambda_i^U \geq 0$ is predictable and does not depend on h.

Range and cap. Before audit t, everything in (18) is known except the stratum drawn and the label revealed. The stratum can be any g with $M_{g,t} > 0$, and the label can be any label in $\mathcal{L}$, including one not yet seen. So the largest possible value is

$$\overline{Z}_t^U = \sum_s \hat{p}_{t-1,s} \psi_{t,s} + \max_{g:\, M_{g,t} > 0} \frac{M_{g,t}}{N \pi_{g,t}} \Big(\max_s \psi_{t,s} - \hat{\mu}_{g,t} \Big). \tag{20}$$

Let $h_t^{\min}$ be the smallest candidate tested at audit t: 0 in the basic version and L_{t-1}^H with cross-capping (proposition 10). We require

$$0 \leq \lambda_t^U \leq \frac{0.9}{\overline{Z}_t^U - h_t^{\min}}, \tag{21}$$

dropping the cap when its denominator is nonpositive.

Proposition 7 (Upper confidence sequence). *Let U_t^H be the smallest $h \geq h_t^{\min}$ with $E_s^U(h) \geq 1/\alpha_+$ for some $s \leq t$, or $U_t^H = \log L$ if there is none, with $h_t^{\min} = 0$. Then $\Pr\{H^\star \leq U_t^H \ \forall t\} \geq 1 - \alpha_+$ for every labeling.*

Proof. Positivity. For $h \geq h_i^{\min}$, the cap (21) gives $1 - \lambda_i^U (Z_i^U - h) \geq 1 - \lambda_i^U (\overline{Z}_i^U - h_i^{\min}) \geq 0.1$.

Supermartingale. At $h = H^\star \geq 0$, lemma 6 gives each factor conditional mean $1 - \lambda_i^U \mathrm{KL}(p^\star \| r_i) \leq 1$. So $E_t^U(H^\star)$ is a nonnegative supermartingale starting at one. By Ville's inequality (lemma 4) it ever reaches $1/\alpha_+$ with probability at most α_+.

Monotonicity. Each factor is positive and nondecreasing in $h \geq 0$. So if some $h < H^\star$ were ruled out at audit s, then $E_s^U(H^\star) \geq E_s^U(h) \geq 1/\alpha_+$.

Start. $\log L \geq H(p)$ for every composition p. □

G.4 THE LOWER END

Smoothing. For any composition p, $1/(N+L) \le \tilde{p}_s \le 1$. At the truth:

- if $p^\star_s > 0$, then $\tilde{p}^\star_s \ge p^\star_s/(1+L/N)$, so $\log(1/\tilde{p}^\star_s) \le \log(1/p^\star_s) + c$ with $c = \log(1+L/N)$;
- if $p^\star_s = 0$, no record is verified as s, so label s never enters $Z_t(p)$.

By lemma 6,

$$\mathbb{E}\big[Z_t(p^\star) \;\big|\; \mathcal{F}_{t-1}\big] \le H^\star + c. \tag{22}$$

On the Kenyan data, $c = \log(1 + 50/33{,}403) = 0.0015$ nats.

Candidates. Every audited record verified as s belongs to $p^\star_s$, so $p^\star \in \mathcal{P}_t = \{p \in \Delta^{L-1} : p_s \ge O_{t,s}/N\}$ for every t, whatever happens in the audit.

The e-processes and the lower end. With a predictable stake $\lambda^L_i \ge 0$ shared by all candidates and all h, let

$$E^L_t(p;h) = \prod_{i \le t} \big\{1 + \lambda^L_i \big(Z_i(p) - h - c\big)\big\}, \qquad F_t(p;h) = \log E^L_t(p;h).$$

At audit t, a value h is ruled out if $\min_{p \in \mathcal{P}_t} F_t(p;h) \ge \log(1/\alpha_-)$, or if a certified lower bound on this minimum is (section G.5). The lower end L^H_t is the largest value ruled out so far, and $L^H_0 = 0$.

How $Z_i(p)$ depends on p. Collecting terms in (19),

$$Z_i(p) = \sum_s \chi_{i,s} \log \frac{1}{\tilde{p}_s} + \varrho_i, \qquad \chi_{i,s} = \frac{1}{N}\Big(O_{i-1,s} + \frac{M_{G_i,i}}{\pi_{G_i,i}} \mathbf{1}\{V_i = s\}\Big) \ge 0, \tag{23}$$

where $\varrho_i = N^{-1}(\sum_g M_{g,i}\hat{\mu}_{g,i} - M_{G_i,i}\hat{\mu}_{G_i,i}/\pi_{G_i,i})$ does not depend on p, and $\sum_s \chi_{i,s} = N^{-1}(i - 1 + M_{G_i,i}/\pi_{G_i,i})$.

The cap. Let $h^{\max}_t$ be the largest candidate tested at audit t: $\log L$ in the basic version and U^H_{t-1} with cross-capping. We require $\lambda^L_t(\sum_s \chi_{t,s} - \varrho_t + h^{\max}_t + c) \le 1$ for every stratum that can be drawn; the label revealed does not affect the left side. Substituting,

$$\lambda^L_t \Big[\frac{t-1}{N} + \frac{M_{g,t}}{N\pi_{g,t}}\big(1+\hat{\mu}_{g,t}\big) - \frac{1}{N}\sum_k M_{k,t}\hat{\mu}_{k,t} + h^{\max}_t + c\Big] \le 1 \quad \text{for every } g \text{ with } M_{g,t} > 0. \tag{24}$$

The bracket is at least $1 + h^{\max}_t + c$. It grows when a large stratum, or one with a large predicted contribution, is drawn with small probability. Priority-aware sampling does exactly this for large, well-known strata, which is why the allocation in section G.6 includes a floor.

Theorem 8 (Lower confidence sequence)**.** *Suppose each λ^L_i is predictable and satisfies (24) with $h^{\max}_i = \log L$. Then $\Pr\{L^H_t \le H^\star \ \forall t\} \ge 1 - \alpha_-$ for every labeling. Together with proposition 7, the interval $[L^H_t, U^H_t]$ satisfies (16).*

Proof. Fix i, $h \le \log L$ and a composition p. Write $\chi_s = \chi_{i,s}$, $\lambda = \lambda^L_i$, $\varsigma = 1 + \lambda(\varrho_i - h - c)$ and $J(p) = \sum_s \chi_s \log(1/\tilde{p}_s) \ge 0$, so that the factor is $\varsigma + \lambda J(p)$.

Positivity. By (24), $\varsigma \ge \lambda \sum_s \chi_s > 0$ when $\lambda > 0$, so every factor is positive.

Monotonicity. Each factor decreases in h, and the stake does not depend on h. So $F_t(p;h)$ is nonincreasing in $h \le \log L$.

Validity. By (22), at $p = p^\star$ and $h = H^\star$ every factor has conditional mean at most one. So $E^L_t(p^\star; H^\star)$ is a nonnegative supermartingale starting at one, and by lemma 4 it ever reaches $1/\alpha_-$ with probability at most α_-.

Suppose $L^H_t > H^\star$. Then some $h \in (H^\star, \log L]$ was ruled out at some audit $s \le t$, by a statistic no larger than $\min_{\mathcal{P}_s} F_s(\cdot; h)$. Since $p^\star \in \mathcal{P}_s$ and F_s is nonincreasing in h, $\log(1/\alpha_-) \le F_s(p^\star; h) \le F_s(p^\star; H^\star)$. This has probability at most α_-.

Two sides. A union bound with proposition 7 gives (16). □

G.5 Computing the lower end: the convex profile

Ruling out h requires the minimum of $F_t(\cdot; h)$ over $\mathcal{P}_t$. The cap that keeps the factors positive also makes this a convex problem.

Proposition 9 (Convex profile)**.** *Under* (24)*, $p \mapsto F_t(p; h)$ is convex on Δ^{L-1} for every $h \le h_t^{\max}$. For any trial point $p^\circ \in \mathcal{P}_t$, the value $\underline{F}_t(h)$ in* (25) *is a lower bound on $\min_{\mathcal{P}_t} F_t(\cdot; h)$. Using it in place of the minimum preserves theorem 8.*

Proof. Keep the notation of the previous proof and let $\upsilon_s(p) = N/(Np_s + 1)$. The derivatives of $\log(1/\tilde{p}_s)$ in p_s are $-\upsilon_s$ and υ_s^2, with no cross terms. For $\zeta(p) = \log\{\varsigma + \lambda J(p)\}$ and any direction y,

$$y^\top \nabla^2 \zeta \, y = \frac{\lambda \sum_s \chi_s \upsilon_s^2 y_s^2}{\varsigma + \lambda J} - \frac{\lambda^2 \big(\sum_s \chi_s \upsilon_s y_s\big)^2}{(\varsigma + \lambda J)^2}.$$

By Cauchy–Schwarz, $\big(\sum_s \chi_s \upsilon_s y_s\big)^2 \le \big(\sum_s \chi_s\big)\big(\sum_s \chi_s \upsilon_s^2 y_s^2\big)$. So this is nonnegative whenever $\varsigma + \lambda J \ge \lambda \sum_s \chi_s$, which the cap gives. Summing over i preserves convexity.

The lower bound is the supporting-plane inequality below. The proof of theorem 8 never uses how the minimum was computed. □

The supporting-plane bound. For a convex differentiable function, $F(p) \ge F(p^\circ) + \nabla F(p^\circ)^\top (p - p^\circ)$ (Vandenberghe & Boyd, 2004). Over $\mathcal{P}_t$, the linear part is smallest with every coordinate at its lower limit $O_{t,s}/N$ and the free mass $1 - t/N$ on the label with the smallest gradient coordinate. Writing $\nabla = \nabla F_t(p^\circ; h)$,

$$\underline{F}_t(h) = F_t(p^\circ; h) - \nabla^\top p^\circ + \frac{1}{N} \sum_s \nabla_s O_{t,s} + \Big(1 - \frac{t}{N}\Big) \min_s \nabla_s \;\le\; \min_{p \in \mathcal{P}_t} F_t(p; h), \tag{25}$$

where

$$\partial_s F_t(p; h) = -\sum_{i \le t} \frac{\lambda_i^L \, \chi_{i,s} \, \upsilon_s(p)}{1 + \lambda_i^L \{Z_i(p) - h - c\}}.$$

In practice. We take p° from projected gradient descent on $\mathcal{P}_t$, with at most 60 iterations, and rule out h when $\underline{F}_t(h) \ge \log(1/\alpha_-)$. The gap $F_t(p^\circ; h) - \underline{F}_t(h)$ is lost evidence, not lost validity, so an inexact optimizer can only widen the interval.

If several stake sequences each satisfy (24), the logarithm of their weighted average, $\log \sum_j w_j \exp F_t^{(j)}$ with $w_j \ge 0$ and $\sum_j w_j = 1$, is again convex in p and nonincreasing in h. This holds because log-sum-exp is convex and increasing in each argument. Mixtures of stake scales are therefore allowed.

G.6 Allocation and cross-capped stakes

Allocation. Strata are submitted label × camera station (300 strata). In our experiment, PAS draws stratum g with probability

$$\pi_{g,t} = 0.4 \frac{M_{g,t}}{\sum_k M_{k,t}} + 0.3 \frac{M_{g,t}(1 + \hat{\mu}_{g,t})}{\sum_k M_{k,t}(1 + \hat{\mu}_{k,t})} + 0.3 \frac{M_{g,t} \hat{\sigma}_{g,t}}{\sum_k M_{k,t} \hat{\sigma}_{k,t}}, \qquad \hat{\sigma}_{g,t}^2 = \sum_s q_{g,t}(s) \big(\psi_{t,s} - \hat{\mu}_{g,t}\big)^2. \tag{26}$$

- *The priority.* The last term is the priority of proposition 3 applied to the influence ψ_t.
- *The floor.* The middle term implies $M_{g,t}(1 + \hat{\mu}_{g,t})/\pi_{g,t} \le \sum_k M_{k,t}(1 + \hat{\mu}_{k,t})/0.3$ for every g, which bounds the stratum-dependent term of (24).
- *The baseline.* SRS sets $\pi_{g,t} = M_{g,t} / \sum_k M_{k,t}$.

Every nonempty stratum has positive probability, so lemma 1 applies.

Proposition 10. *Run both ends on the same audits with $h_t^{\min} = L_{t-1}^H$ in* (21) *and $h_t^{\max} = U_{t-1}^H$ in* (24)*, testing only candidates in $[L_{t-1}^H, U_{t-1}^H]$. Then* (16) *holds.*

Proof. The endpoints only move inward, and they start at $0 \le H^\star \le \log L$. Let t_0 be the first audit after which $H^\star \notin [L^H_{t_0}, U^H_{t_0}]$. The modified stakes

$$\lambda_i^{U\prime} = \lambda_i^U \mathbf{1}\{H^\star \ge L^H_{i-1}\}, \qquad \lambda_i^{L\prime} = \lambda_i^L \mathbf{1}\{H^\star \le U^H_{i-1}\}$$

are predictable, since $H^\star$ is a fixed number. With them, the factors at the truth are positive and have conditional mean at most one, so both modified processes are nonnegative supermartingales. They reach $1/\alpha_+$ and $1/\alpha_-$ with probabilities at most α_+ and α_-.

For $i \le t_0$, $H^\star \in [L^H_{i-1}, U^H_{i-1}]$, so the modified stakes equal the actual ones up to audit t_0. A failure at t_0 is one of two events:

- some $h \in [L^H_{t_0-1}, H^\star)$ has $E^U_{t_0}(h) \ge 1/\alpha_+$, which forces $E^U_{t_0}(H^\star) \ge 1/\alpha_+$;
- some $h \in (H^\star, U^H_{t_0-1}]$ is ruled out by the lower end, which forces $F_{t_0}(p^\star; H^\star) \ge \log(1/\alpha_-)$.

Either is a crossing of a modified process, and a union bound finishes the proof. □

The proof uses only that the ceiling $h_t^{\max}$ is predictable and never increases, not that it is an upper confidence bound. So a one-sided lower bound can use any such ceiling, for example one set near the current plug-in entropy.

H SCORES BEYOND THE BIODIVERSITY PILLARS

Ratios. For $g(C) = \langle U, C\rangle / \langle W, C\rangle$ with $W \ge 0$, $\rho^\star = g(C^\star)$ satisfies $\langle U - \rho^\star W, C^\star\rangle = 0$, so the observation with contributions $U - \rho W$ has conditional mean zero at $\rho = \rho^\star$ (lemma 1). An e-process for each candidate ρ gives a confidence sequence for recall, F_1 and balanced accuracy.

Bounded curvature. If $\nabla^2 g$ is bounded by β_g on the relevant region, then $g(p^\star) = g(r) + \nabla g(r)^\top (p^\star - r) + R$ with $|R| \le \frac{\beta_g}{2}\|p^\star - r\|^2$; a linear confidence sequence for the first-order term and a composition region for the remainder, at a separate budget, certify g. Shannon entropy is not of this type, because $\partial^2 H / \partial p_s^2 = -1/p_s$ is unbounded near zero.

Support-based scores. Richness changes by one when a single record reveals a new species. If a procedure declares an eligible species absent after T reviews and errs with probability at most α under every alternative in which one record carries it, then $\mathbb{E}[T] \ge (1-\alpha)N$: couple with each single-record alternative, which agrees until that record is reviewed, and sum over records. Presence, by contrast, is certified by one verified record.