When Agreement Labels Are Necessary for Worst-Group Model Selection

Anonymous authors
Paper under double-blind review

Abstract

Labels on points where all candidate classifiers agree cancel from average-risk comparisons, yet can determine the minimizer of worst-group error. We characterize when this agreement information is necessary for selecting among fixed binary classifiers. On a finite audit pool, a closed-form certificate gives the exact largest regret over a product of label-count intervals and supports adaptive querying and stopping through finite-population confidence sequences. Once all disagreement labels are known, an $O(MG)$ criterion determines whether a candidate meets the regret tolerance under every completion of the agreement labels. An indistinguishability bound establishes when no uniformly reliable agreement-free selector exists. Matching oracle bounds show that a second group can eliminate the quadratic label savings of small disagreement mass. CAWS uses the certificate with disagreement-first sampling and agreement fallback. Fixed synthetic, digit, Adult, and CelebA pools illustrate both sides of the criterion: four of ten CelebA banks exhibit an agreement obstruction. Matched-path ablations isolate the reduction in certification cost obtained by preserving shared losses.

1 Introduction

Worst-group validation is used to select checkpoints and hyperparameters when average error masks poor performance on a small group (Sagawa et al., 2020; Kirichenko et al., 2023; Yang et al., 2023). Because labels can be expensive, it is attractive to query only points where the candidates disagree. For average-risk comparison, this disagreement sampling is easy to justify: when two candidates make the same prediction, their losses coincide for either label, so the point contributes nothing to their risk difference (Sawade et al., 2012; Hara et al., 2024; Kay et al., 2025).

Worst-group selection breaks this argument. A shared error raises one group's risk for every candidate, but that group need not attain every candidate's maximum. The label of an agreement point can therefore change the preferred candidate even when every within-group risk difference is known, and a learner restricted to disagreement labels cannot distinguish such labelings. Figure 1 shows how CAWS combines disagreement-first sampling with a certificate that retains group risk levels.

Our central question is when agreement labels can be omitted while still guaranteeing a near-optimal choice. We answer it for a fixed audit pool partitioned by observed group and joint prediction pattern. Each queried label reveals every candidate's error on that point. For binary predictions, this shared observation yields an exact, closed-form expression for the largest selection regret over a box of label-count intervals (Theorem 1). The interval construction determines statistical precision; the closed form introduces no further optimization slack within that box.

Combined with anytime-valid count intervals, the expression certifies selection under adaptive acquisition and stopping. After a disagreement census, which labels every disagreement point, it gives a necessary-and-sufficient criterion for a candidate to be near-optimal under *every* completion of the agreement labels (Proposition 1). If the criterion fails, no agreement-free selector that always returns a candidate is uniformly δ-correct for $\delta < 1/\min\{M, G\}$. Separate oracle bounds characterize intrinsic label requirements over all valid selectors: the minimax query complexity grows from $\Theta(q^2\tau^{-2}\log(1/\delta))$ with one group to $\Theta(\tau^{-2}\log(1/\delta))$ with two, where q bounds the within-group disagreement mass.

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.

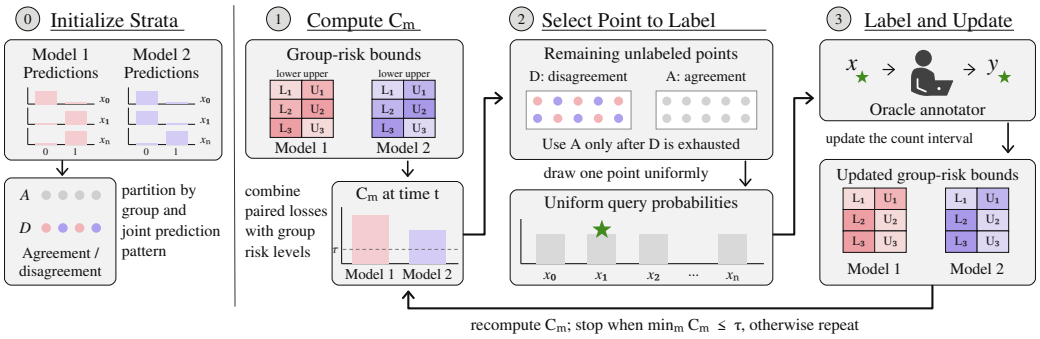


Figure 1: **CAWS workflow.** Groups and joint prediction patterns define the strata. Each label updates a count interval and the shared-loss certificate. Until some candidate satisfies $C_m \leq \tau$, CAWS samples uniformly from the remaining disagreement points and, once these are exhausted, from the remaining agreement points. Matrices and bars are schematic.

We use **CAWS** (Certified Agreement-Aware Worst-Group Selection) to instantiate the certificate with a simple acquisition rule: query disagreement points first, then query agreement points if certification remains unresolved at the disagreement census. The experimental study examines the completion criterion on fixed synthetic, digit, Adult, and CelebA prediction pools. Four of the ten CelebA banks exhibit agreement obstructions, while the digit and Adult pools admit agreement-free certification. Comparisons under a common certificate measure acquisition cost, and matched-path ablations isolate the contribution of preserving shared losses. Together, these analyses connect agreement necessity, computable certification, and the labeling cost of a certified choice.

2 Related Work

Selecting models with limited labels. Active comparison (Sawade et al., 2012), variance-based selection (Hara et al., 2024), Model Selector (Okanovic et al., 2025), and CODA (Consensus-Driven Active Model Selection) (Kay et al., 2025) choose informative queries or exploit dependence among pretrained classifiers. Model Selector scores unlabeled points by their expected value for identifying the best candidate, whereas CODA models classifier–class–point relations and uses posterior consensus to prioritize labels. SELECT-LLM extends information-driven selection to generative outputs through response similarities (Durmazkeser et al., 2026). Active testing targets label-efficient risk estimation (Kossen et al., 2021), and active Bayesian and attribute-level assessment estimate classifier performance, including on subgroups (Ji et al., 2021; Piratla et al., 2021). Active fairness auditing also selects informative labels (Yan and Zhang, 2022), recently for black-box LLMs with surrogate version spaces (Hartmann et al., 2026). We study the information needed to select under a maximum of observed group risks. Agreement labels can change this objective even though they cancel from every within-group difference; our completion criterion identifies exactly when a common near-optimal choice survives that uncertainty.

Sequential and robust selection. Confidence sequences support anytime-valid inference under optional stopping (Howard et al., 2021; Waudby-Smith and Ramdas, 2020); prediction-powered and active inference use predictions to reduce labeling cost (Angelopoulos et al., 2023; Zrnic and Candès, 2024). We construct the count region from finite-population mixture sequences. Maximin action identification samples action-pair rewards (Garivier et al., 2016), covariance-adaptive identification exploits dependent observations (Saad et al., 2023), and distributionally robust selection compares alternatives across input distributions (Fan et al., 2020). In our setting, each query reveals a full loss vector at a known group and prediction pattern, and labels are sampled without replacement. Binary loss compatibility makes the regret bound attainable over the count box. Sequential change of measure (Kaufmann et al., 2016) then locates the information required for selection in agreement strata, where every within-group loss difference is zero.

3 SETTING: A CERTIFIED DECISION ON A FIXED POOL

The audit pool has N points in G nonempty disjoint groups, together with predictions from $M \geq 2$ binary classifiers. The candidates and groups are fixed before acquisition, and the candidates are fitted on training records disjoint from the audit pool. Acquisition observes the predictions, the group memberships, and the labels it requests. A group may be a recorded environment or demographic attribute available before the target label is queried.

Partition each group into strata by the full prediction vector. Each nonempty stratum s has size N_s, group $g(s)$, candidate predictions $v_{sm} \in \{0,1\}$, and within-group mass $p_s = N_s/N_{g(s)}$, where N_g is the size of group g. Its unknown positive-label fraction is $\theta_s = K_s/N_s$, and $\theta^\star$ denotes the vector of these true fractions. Candidate m's group risks, worst-group risk, and regret are

$$\begin{aligned} R_{mg}(\theta) &= \sum_{s:g(s)=g} p_s\big[v_{sm} + (1-2v_{sm})\theta_s\big], \\ W_m(\theta) &= \max_g R_{mg}(\theta), \qquad \mathrm{Reg}_m(\theta) = W_m(\theta) - \min_j W_j(\theta). \end{aligned} \tag{1}$$

At each round, the learner chooses a stratum and queries a uniformly drawn unqueried point in it; the label reveals every candidate's loss on that point. For a prespecified tolerance $\tau \geq 0$ and failure level $\delta \in (0,1)$, we require

$$\mathbb{P}\big\{\exists t:\ \text{a candidate } \widehat{m}_t \text{ is certified and } \mathrm{Reg}_{\widehat{m}_t}(\theta^\star) > \tau\big\} \leq \delta. \tag{2}$$

This is a finite-pool statement over query randomization, conditional on the fixed labels.

After n_s queries in stratum s with k_s positives, we obtain $[l_s, u_s]$ by inverting a finite-population mixture process: for each feasible total K, let

$$E_{s,n_s}(K) = \frac{\binom{N_s}{n_s}}{(n_s+1)\binom{K}{k_s}\binom{N_s-K}{n_s-k_s}}, \qquad \mathcal{K}_s = \{K : E_{s,n_s}(K) \leq \alpha_s^{-1}\}. \tag{3}$$

The interval endpoints are $\min \mathcal{K}_s/N_s$ and $\max \mathcal{K}_s/N_s$. Under the true count, E_{s,n_s} is a nonnegative supermartingale, so Ville's inequality gives coverage simultaneously at all query counts in the stratum. We allocate $\alpha_s = \delta p_s/G$ using metadata only. Since $\sum_s \alpha_s = \delta$, a union bound shows that the box $\mathcal{B}_t = \prod_s [l_s(t), u_s(t)]$ contains $\theta^\star$ at every round with probability at least $1-\delta$, under any adaptive interleaving of strata (Appendix A). A census, which queries every point of a stratum, collapses its interval to a point. Primary experiments use current intervals; running intersections are evaluated separately on unchanged query paths.

4 EXACT REGRET AND AGREEMENT COMPLETION

4.1 A CERTIFICATE THAT PRESERVES SHARED LOSSES

For a nonempty box $\mathcal{B}_t$, define

$$L_{mg} = \min_{\theta \in \mathcal{B}_t} R_{mg}(\theta), \quad U_{mg} = \max_{\theta \in \mathcal{B}_t} R_{mg}(\theta), \quad B_{mjg} = \max_{\theta \in \mathcal{B}_t} \big(R_{mg}(\theta) - R_{jg}(\theta)\big). \tag{4}$$

All three quantities are endpoint sums:

$$\begin{aligned} L_{mg} &= \sum_{s:g(s)=g} p_s[(1-v_{sm})l_s + v_{sm}(1-u_s)], \\ U_{mg} &= \sum_{s:g(s)=g} p_s[(1-v_{sm})u_s + v_{sm}(1-l_s)], \\ B_{mjg} &= \sum_{s:g(s)=g} p_s\big[(1-v_{sm})v_{sj}(2u_s-1) + v_{sm}(1-v_{sj})(1-2l_s)\big]. \end{aligned} \tag{5}$$

Agreement coordinates vanish from B_{mjg}, which may be negative and must not be clipped at zero. Let $\underline{W}_j = \max_h L_{jh}$, a lower bound on candidate j's worst-group risk.

Theorem 1 (Exact binary-box regret)**.** *For any nonempty product box of binary label fractions, set*

$$\begin{aligned} C_{mj} &= \max_g \min\{B_{mjg},\ U_{mg} - \underline{W}_j\}, \quad j \neq m, \\ C_{mm} &= 0, \qquad C_m = \max_j C_{mj} = \sup_{\theta \in \mathcal{B}_t} \mathrm{Reg}_m(\theta). \end{aligned} \tag{6}$$

For the simultaneous count box above, with probability at least $1-\delta$, $\mathrm{Reg}_m(\theta^\star) \le C_m(t)$ for every candidate m and round t. Certifying a candidate only when $C_m(t) \le \tau$ therefore guarantees (2).

Why the expression is exact. For a candidate–competitor–group triple (m, j, g), $R_{mg} - W_j$ is at most both the within-group difference B_{mjg} (since $W_j \ge R_{jg}$) and the level bound $U_{mg} - \underline{W}_j$ (since $W_j \ge \underline{W}_j$). Crucially, a single corner attains the smaller of the two. In group g, set $\theta_s = u_s$ when $v_{sm} = 0$ and $\theta_s = l_s$ when $v_{sm} = 1$. This maximizes both candidate m's risk and its difference from j: coordinates where m and j agree drop out of the difference, and the others enter both objectives with the same sign. In every other group, independently minimize j's risk. Maximizing over g and j proves the formula, since $W_m = \max_g R_{mg}$.

For the count intervals above, every attaining endpoint is an accepted integer count divided by its stratum size. The maximizing corner is therefore a feasible labeling of the unqueried points, and optimizing over the discrete acceptance sets gives the same value as optimizing over their interval hulls. After interval inversion, evaluating all candidates costs $O(SM^2 + GM^2)$ for S nonempty strata.

4.2 When can all agreement labels be omitted?

Within group g, let D_g contain the points on which not all candidates agree, and let A_g contain the remaining agreement points. Suppose all disagreement labels are revealed. Their contribution to candidate m's risk is

$$d_{mg} = \frac{1}{N_g} \sum_{i \in D_g} \mathbf{1}\{f_m(x_i) \neq y_i\}, \qquad a_g = \frac{|A_g|}{N_g}.$$

We call each labeling of the agreement points a *completion*. Every completion gives risks $R_{mg} = d_{mg} + c_g$, where c_g is the shared agreement-error contribution, whose feasible set is exactly $\mathcal{C}_A = \prod_g \{0, 1/N_g, \dots, a_g\}$.

Proposition 1 (Agreement-completion criterion)**.** *The largest regret of candidate m over all completions is*

$$\begin{aligned} C_m^D &= \max_{c \in \mathcal{C}_A} \mathrm{Reg}_m(d + c) \\ &= \max_{j,g} \min\left\{ d_{mg} - d_{jg},\ d_{mg} + a_g - \max_h d_{jh} \right\}, \qquad \rho_D = \min_m C_m^D. \end{aligned} \tag{7}$$

Some candidate is τ-optimal under every completion if and only if $\rho_D \le \tau$. Each C_m^D is attained by a completion in which every agreement prediction is wrong in one group and correct in all others. If $\rho_D > \tau$, every randomized selector that observes no agreement labels and almost surely returns a candidate has error probability at least $1/\min\{M, G\}$ on some completion.

The formula includes $j = m$, making the regret nonnegative. It shows that ruling out m requires two conditions in the same group: the paired disadvantage there must exceed the tolerance, and the attainable risk level there must be high enough to matter against the competitor's maximum.

To see the randomized obstruction, reveal all disagreement labels for free; the selector's output distribution is then the same under every completion. If $\rho_D > \tau$, every candidate fails on at least one of the G endpoint completions above. Summed over these completions, the failure probabilities total at least one, so some completion has error probability at least $1/G$. The most likely output also fails on some completion, giving $1/M$. With two groups, no agreement-free selector thus succeeds uniformly with probability above $1/2$ unless a robust candidate exists, regardless of the number of candidates.

Before the disagreement census, applying Theorem 1 with agreement coordinates left at $[0, 1]$ gives an anytime-valid, agreement-free certificate that equals (7) at the census. A complementary

sufficient condition is a common resolved worst group: if $L_{mg_*} \geq \max_{h \neq g_*} U_{mh}$ for every m, then $C_m = \max_j B_{mjg_*}$ (Proposition 3). Such a common worst group suffices for agreement labels to cancel but is not necessary: the completion criterion also covers pools whose candidates have different worst groups across feasible labelings.

An $O(MG)$ completion test. The endpoint completions give a faster way to evaluate (7). For each group g, let $c^{[g]} = a_g e_g$, with e_g the gth basis vector, and define

$$w_m^{[g]} = \max\left\{ d_{mg} + a_g,\ \max_{h \neq g} d_{mh} \right\}, \quad b_g = \min_j w_j^{[g]}, \quad C_m^D = \max_g \left(w_m^{[g]} - b_g \right). \tag{8}$$

Here $w_m^{[g]}$ is candidate m's worst-group risk under completion $c^{[g]}$. By Proposition 1, one of these G completions attains each candidate's maximum regret, which proves the last equality. The two largest entries in each row of d supply every leave-one-group maximum in constant time; for $G = 1$, that maximum is $-\infty$. Computing all $w_m^{[g]}$, b_g, and C_m^D therefore takes $O(MG)$ time once d and a are available, instead of a scan over all candidate pairs.

A completion obstruction in two groups. In the two Level-trap pools (Section 6.3),

$$d = \begin{pmatrix} 0 & 0.2 \\ 0.2 & 0 \end{pmatrix}, \qquad a = (0.8, 0.8).$$

At $c^{[1]} = (0.8, 0)$, the two worst-group risks are $(0.8, 1)$, giving regrets $(0, 0.2)$; at $c^{[2]} = (0, 0.8)$, they are $(1, 0.8)$, giving regrets $(0.2, 0)$. Hence $C_1^D = C_2^D = \rho_D = 0.2$, a value determined by the disagreement labels and metadata alone. The two pools differ only in their agreement-error counts, $(60, 100)$ and $(140, 100)$, which yield different winners, each with gap 0.1; the same completion obstruction applies to both. For one group, by contrast, adding c_1 shifts every risk equally, so $C_m^D = d_{m1} - \min_j d_{j1}$ and $\rho_D = 0$.

4.3 CAWS: CERTIFICATION WITH AGREEMENT FALLBACK

Algorithm 1: CAWS

Input: fixed predictions, groups, tolerance τ, failure level δ. Initialize strata, counts, and finite-population intervals from metadata.

1. Compute C_m for every candidate using (6).
2. If some $C_m \leq \tau$, return the eligible candidate with the smallest $\widehat{W}_m$; break remaining ties by model index.
3. Sample uniformly from the unqueried disagreement points. If none remain, sample uniformly from the remaining pool.
4. Query the label, update its stratum's count interval, and return to step 1.

Here $\widehat{W}_m$ uses a uniform working prior on each stratum's total positive count, with posterior mean fraction

$$\bar{\theta}_s = \frac{k_s + (N_s - n_s)(k_s + 1)/(n_s + 2)}{N_s}, \qquad \widehat{W}_m = \max_g R_{mg}(\bar{\theta}). \tag{9}$$

The certificate determines eligibility, $\widehat{W}_m$ chooses among eligible candidates, and model index breaks remaining ties. At a budget checkpoint before certification, the primary experiments recommend the smallest C_m, breaking ties by $\widehat{W}_m$ and then model index.

Corollary 1 (Termination at the disagreement census)**.** *With current count intervals, if $\rho_D \leq \tau$, CAWS certifies after at most $\sum_g |D_g|$ labels and queries no agreement point. On every pool, it certifies after at most N labels.*

Proof. After the last disagreement query, all disagreement intervals are singletons and the untouched agreement intervals are $[0, 1]$. Thus $\min_m C_m = \rho_D \leq \tau$, and the stopping condition holds. At a full census the box is a singleton, so a true minimizer has certificate zero. □

5 How costly can agreement information be?

We quantify the intrinsic query complexity in an oracle with known group–prediction masses, where each query to a positive-mass stratum returns a fresh independent label. With two candidates and at most q disagreement mass in each group, let $T^*_G(q, \tau, \delta)$ be the minimax expected query count over all selectors that return a τ-optimal candidate with probability at least $1 - \delta$ on every instance.

Theorem 2 (One-group/two-group separation). *For $0 < q \le 1/4$, $0 < \tau \le q/4$, and $0 < \delta \le 1/10$,*

$$T^*_1 = \Theta\left(\frac{q^2}{\tau^2} \log \frac{1}{\delta}\right), \qquad T^*_2 = \Theta\left(\frac{1}{\tau^2} \log \frac{1}{\delta}\right). \tag{10}$$

The first upper bound uses disagreement labels only, and uniform within-group sampling attains the second. For a pair of two-group instances, every uniformly valid selector requires at least $\mathrm{kl}(1-\delta, \delta)/(512\tau^2)$ agreement queries in expectation on one of them.

With one group, a disagreement label yields a signed loss difference $Z \in \{-1, 1\}$, and the risk difference is $d\,\mathbb{E}Z$ for a known $d \le q$. Deciding the τ-optimal candidate thus costs $O(q^2\tau^{-2}\log(1/\delta))$ queries; a matching lower bound perturbs only the disagreement-label distribution.

With two groups, use agreement prediction $(0, 0)$ with mass $1 - q$ in both groups, disagreement predictions $(0, 1)$ in group 1 and $(1, 0)$ in group 2, and zero disagreement labels. Setting the shared agreement errors to $a \pm h$ and a in the two groups gives the candidate-by-group risk matrices

$$R^{\pm}(h) = \begin{pmatrix} a \pm h & a + q \\ a \pm h + q & a \end{pmatrix}, \qquad a = \tfrac{1}{4}, \quad h = 2\tau \le q/2. \tag{11}$$

The optimal candidate switches, with gap $h > \tau$, yet only group 1's agreement-label distribution changes. Each such agreement query carries $O(\tau^2)$ KL divergence and every other query carries none, so adaptive change of measure gives the lower bound. Estimating the four group risks by uniform within-group sampling gives the matching upper bound. The information that destroys the q^2 savings thus resides in agreement labels.

For a finite pool (Proposition 2), retain $0 < q \le 1/4$ and $0 < \delta \le 1/10$, and use the same prediction patterns in two groups of n points with qn integral. If $0 < \tau \le q/8$ and $(1 - q)n \ge 2\tau^{-2}\log(4/\delta)$, every terminating selector that is δ-correct on all labelings has a fixed pool on which its number $N_A(T)$ of agreement queries satisfies

$$\mathbb{E}N_A(T) \ge \frac{\mathrm{kl}(1 - 2\delta, 2\delta)}{2048\tau^2}. \tag{12}$$

The proof draws independent agreement labels with shift $h = 4\tau$. Concentration preserves the ranking and leaves a finite-pool gap of at least 3τ; change of measure then forces the stated number of agreement queries. The size condition ensures that concentration takes effect before the census cap becomes binding. The completion criterion, by contrast, applies at every pool size.

6 Experiments

We evaluate certification cost, identify pools that require agreement labels, and isolate the contribution of the shared-loss certificate.

Protocol and metrics. Candidates are frozen before acquisition. All policies use the same finite-population intervals, error allocation, exact certificate, and tolerance. We report the certification rate (CR) and restricted mean time to certification (RMTC), in labels, at label cap B:

$$\mathrm{CR}(B) = \mathbb{P}(T \le B), \qquad \mathrm{RMTC}(B) = \mathbb{E}[\min(T, B)], \tag{13}$$

where T is the number of labels at the first issued certificate, and $T = \infty$ if a run ends without one. Abstentions and executions ending without a certificate contribute B to RMTC. CR measures certificate issuance, not correctness: an incorrect certificate also counts toward CR. A completed census followed by certificate evaluation always permits certification.

The primary cap is the audit-pool size, and $\delta = 0.05$. The five synthetic pools (Appendix G.1) use 40 query repeats each, with $\tau = 0.01$ for Rare and 0.04 otherwise. Broad and Local digits (scikit-learn

Table 1: RMTC (labels) on digits and Adult; every entry has CR $= 100\%$. Standard errors use five fitted-pool means, with 25 query repeats for digits and ten for Adult.

Policy	Broad	Local	Adult
Uniform	224.5 ± 15.4	56.2 ± 10.1	686.0 ± 19.4
Variance (Hara et al., 2024)	206.4 ± 12.7	255.7 ± 13.3	712.4 ± 16.5
Bayes-IG (Ji et al., 2021)	175.7 ± 7.5	12.3 ± 2.7	135.6 ± 6.6
Disagreement (Sawade et al., 2012)	116.8 ± 7.3	6.1 ± 1.7	103.9 ± 5.6
Mix10	122.4 ± 7.8	6.8 ± 2.1	115.3 ± 5.3
WAVE lookahead	170.0 ± 13.6	14.3 ± 10.3	273.5 ± 29.1
CAWS (ours)	116.1 ± 7.0	5.6 ± 1.8	104.9 ± 5.0

developers, 2026) have five training seeds, 12 logistic-regression candidates, 719 audit points, two observable nuisance groups, and 25 repeats per seed; $\tau = 0.05$. Adult (Becker and Kohavi, 1996) has five splits, 32 logistic-regression candidates, 1,000 audit points, and ten repeats per split; $\tau = 0.02$, and the recorded sex attribute defines the groups. Each split fits candidates on 31,561 records; the official test set is excluded from fitting, querying, selection, and stopping. Standard errors first average query repeats within each fitted pool, then use the five seed or split means.

We compare Uniform, Variance, Bayes-IG, Disagreement, CAWS, Mix10, and WAVE (Appendix F). Mix10 samples uniformly from the whole pool with probability 0.1 while disagreement points remain; WAVE is a one-step lookahead control that minimizes the expected next value of $\min_m C_m$. The CelebA study uses fixed deep-model banks to examine the agreement-completion criterion and the resulting query behavior.

6.1 Certification on digits and Adult

All seven policies certify in every run on these pools (Table 1). CAWS uses 116.1 labels on Broad and 104.9 on Adult, compared with 224.5 and 686.0 for Uniform. Every digit and Adult pool satisfies $\rho_D \leq \tau$, so CAWS certifies without agreement queries. Its query distribution coincides with that of Disagreement until exhaustion; small cost differences reflect query randomization.

Local begins close to the requested tolerance: its initial minimum certificate is 0.053–0.059, and 7–11 of its 12 candidates are within $\tau = 0.05$ of optimal. In a separate sensitivity study with three repeats per seed, reducing τ from 0.05 to 0.02 and 0.01 increases the cost of CAWS from 5.1 to 28.2 and 36.2 labels (Table 18). The low cost at the primary tolerance thus reflects the ease of the certified decision.

Adult also permits an independent test evaluation. The held-out worst-group error of the CAWS selection is 0.187, compared with 0.188 for Uniform and Disagreement. Within each split, all 32 candidates share a realized worst group, and 16–27 meet the audit tolerance. The benefit here is a lower auditing cost at similar test error. The certificate concerns the audit pool, whereas held-out error measures the selected candidate on records excluded from acquisition.

6.2 Agreement completion in fixed deep-model banks

On CelebA (Liu et al., 2015), the target is `Blond_Hair` and the two groups are defined by `Male`. Each audit pool contains 4,000 held-out images and 16 candidates spanning ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2021), standard and group-balanced training, regularization strengths, and checkpoints. The study comprises ten separately trained candidate banks, with 25 query repeats per bank at $\tau = 0.02$ and $\delta = 0.05$.

The completion diagnostic separates these fixed banks into four obstructed and six unobstructed instances (Table 2). On the four obstructed banks, $\rho_D \in [0.031, 0.044] > \tau$. Since $G = 2$, Proposition 1 implies that any agreement-free selector that always returns a candidate has error probability at least $1/2$ on some completion of each such bank. This statement is determined by the disagreement labels and prediction metadata.

Disagreement abstains on the four obstructed banks and certifies on the other six. CAWS certifies in every run, querying agreement labels on precisely the four obstructed banks. Its mean agreement-query count is 711.5 ± 36.2 on those banks and zero on the remaining six. Thus both branches of the

Table 2: Agreement completion on CelebA at $\tau = 0.02$. Each bank has 25 query repeats. CR denotes certificate issuance. Agreement-query counts are for CAWS; standard errors use four bank means for the obstructed subset and ten overall.

Bank subset	Banks	ρ_D	D-only CR	CAWS CR	Agreement
Obstructed	4	0.031–0.044	0%	100%	711.5 ± 36.2
Unobstructed	6	≤ 0.020	100%	100%	0
All banks	10	—	60%	100%	284.6 ± 116.9

Table 3: Synthetic RMTC (labels) over 40 query repeats. The two † entries have CR = 0% and are charged the 800-label cap; all other entries have CR = 100%. Table 6 gives standard errors.

Policy	Rare	One group	Level +	Level −	Diverse
Uniform	333.4	103.6	473.9	455.1	557.7
Disagreement (Sawade et al., 2012)	13.8	20.6	800†	800†	540.2
WAVE lookahead	13.8	20.6	474.9	469.5	533.8
CAWS (ours)	13.8	20.6	472.8	465.5	543.9

completion criterion occur among fixed deep classifiers on the same visual task. Appendix L reports the associated query costs.

6.3 Controlled agreement obstructions

Table 4 summarizes completion feasibility. Among the 20 synthetic, digit, and Adult pools, only the two Level traps have $\rho_D > \tau$, both with $\rho_D = 0.20$ at $\tau = 0.04$. Disagreement exhausts their 160 disagreement points and abstains, whereas CAWS continues to certification after 472.8 and 465.5 labels, including 312.8 and 305.5 agreement queries (Table 3); Uniform has similar costs. On Rare and One group, where agreement-free certification is possible, CAWS uses 13.8 and 20.6 labels versus 333.4 and 103.6 for Uniform. Diverse also permits agreement-free certification but requires nearly a census: feasibility alone does not imply inexpensive certification.

6.4 Matched-path certificate ablation

Holding queried labels and stratum intervals fixed, we compare the exact certificate with a marginal-only certificate that bounds each candidate's risk separately (Appendix I). On Broad WAVE paths, preserving shared-loss constraints reduces mean cost from 644.1 to 174.8 labels; on Local, from 665.3 to 14.1 (Table 5). On Level trap +, the costs are equal because uncertainty in shared risk levels still determines stopping.

An exact certificate alone does not ensure efficient acquisition. On Local seed 29, WAVE spends 47.8 agreement labels per run in rounds where all actions have zero predicted gain and ties are broken by interval width. In these rounds, the binding certificate component is a paired difference, in which agreement coefficients vanish. This explains how one-step lookahead can spend labels without relaxing the constraint that prevents certification.

Mechanism and prediction fragmentation. An agreement-level sweep uses 20 settings with two groups of 800 points, varying the disagreement mass q and the shared-risk shift h, with $\tau = q/8$; each setting has 20 query repeats under Uniform, Disagreement, CAWS, and WAVE, for 1,600 trajectories in total. Disagreement abstains in every setting, while the other three policies always certify. At $q = 0.20$, WAVE's mean cost increases from 577.6 labels at $h/q = 0.75$ to 1,442.2 at $h = 0$. Even when the candidates tie on the realized pool, certifying a small regret requires ruling out agreement completions that favor either candidate. Early agreement sampling can also reduce cost: at $q = 0.20$ and $h/q = 0.75$, Uniform uses 467.9 labels and CAWS uses 559.4. The need for agreement information therefore does not prescribe a single best query order.

A second control replaces prediction-stratum intervals with group-pooled loss and paired-indicator confidence sequences, allocating error across GM^2 indicators and holding uniform point orders

Table 4: Agreement-completion diagnostic. A robust agreement-free choice exists exactly when $\rho_D \leq \tau$. The 20 synthetic, digit, and Adult pools and the ten CelebA banks cover both sides of this criterion.

Pool family	Pools	τ	ρ_D	Feasible
Rare disagreement	1	0.01	0	1/1
One group	1	0.04	0	1/1
Level trap +	1	0.04	0.200	0/1
Level trap −	1	0.04	0.200	0/1
Diverse	1	0.04	0.017	1/1
Broad digits	5	0.05	0–0.006	5/5
Local digits	5	0.05	0	5/5
Adult	5	0.02	0–0.006	5/5
CelebA, unobstructed	6	0.02	≤ 0.020	6/6
CelebA, obstructed	4	0.02	0.031–0.044	0/4

Table 5: Matched-path certificate ablation: labels to certification for the exact and marginal-only certificates on identical observations and intervals. Synthetic errors use ten query repeats; digit errors use five seed means with ten repeats each.

Pool	Query path	Exact	Marginal only
Rare	WAVE	14.5 ± 1.0	957.8 ± 5.3
Level trap +	WAVE	461.3 ± 22.6	461.3 ± 22.6
Broad	Uniform	225.6 ± 17.0	665.3 ± 1.4
Broad	WAVE	174.8 ± 11.1	644.1 ± 4.1
Local	WAVE	14.1 ± 10.3	665.3 ± 1.0

fixed. At $N = 1{,}024$, $M = 64$, and $\tau = 0.05$, nearly every prediction stratum is a singleton. For $G = 2, 4, 8$, the count box requires 81.0%, 88.7%, and 93.1% of labels, versus 87.8%, 93.1%, and 95.0% for the pooled control. Both constructions incur substantial labeling cost when prediction strata have little repeated structure.

Reliability and computation. Across the seven policies on the synthetic, digit, and Adult pools, 3,500 runs yield 3,420 certificates and 80 abstentions, and no issued certificate is incorrect. One trajectory has an intermediate failure of the regret bound; we track such confidence-bound failures separately from failures of the certified decision. Independent checks compare the certificate with linear programs and enumerate finite-population coverage (Appendix H). At $M = 64$, $G = 8$, and $N = 1{,}024$, a certificate evaluation takes 39.87 ms, versus 2,188.37 ms for WAVE's batched lookahead, which evaluates both label outcomes in every available stratum.

7 Discussion

Agreement labels determine group risk levels even when all within-group differences are known. The completion criterion identifies exactly when this unresolved information prevents a common near-optimal choice, and the oracle separation quantifies its possible label cost. The binary-box certificate makes this structure usable during adaptive acquisition, with exact optimization over the chosen count region. CAWS combines that certificate with disagreement-first sampling and agreement fallback. The fixed-pool experiments exhibit both agreement-free and obstructed decisions, while matched-path ablations show how retaining shared losses can reduce the evidence needed for certification.

Limitations. The theory concerns fixed binary candidates and a finite audit pool. With candidates independent of within-group audit sampling, simultaneous pool-to-population risk error at most ϵ yields population regret at most $\tau + 2\epsilon$ (Proposition 4). Certification can approach a census in fragmented strata. Instance-adaptive acquisition, extensions to other losses, and inference that pools information across strata remain open directions.

AI USE STATEMENT

Generative AI tools assisted with language editing, manuscript organization, critical feedback on the presentation of methods and experimental results, and preparation of the schematic figure. The authors take full responsibility for the final manuscript, including its theoretical claims, proofs, experimental results, references, and figures.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382:669–674, 2023. https://doi.org/10.1126/science.adi6000.

Barry Becker and Ronny Kohavi. Adult [Dataset]. *UCI Machine Learning Repository*, 1996. https://doi.org/10.24432/C5XW20.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. https://arxiv.org/abs/2010.11929.

Yavuz Durmazkeser, Patrik Okanovic, Andreas Kirsch, Torsten Hoefler, and Nezihe Merve Gürel. Large language model selection with limited annotations. arXiv:2605.24981, 2026. https://arxiv.org/abs/2605.24981.

Weiwei Fan, L. Jeff Hong, and Xiaowei Zhang. Distributionally robust selection of the best. *Management Science*, 66(1):190–208, 2020. https://arxiv.org/abs/1903.05828.

Aurélien Garivier, Emilie Kaufmann, and Wouter M. Koolen. Maximin action identification: A new bandit framework for games. In *Conference on Learning Theory*, PMLR 49, pp. 1028–1050, 2016. https://proceedings.mlr.press/v49/garivier16b.html.

Satoshi Hara, Mitsuru Matsuura, Junya Honda, and Shinji Ito. Active model selection: A variance minimization approach. *Machine Learning*, 113:8327–8345, 2024. https://doi.org/10.1007/s10994-024-06603-1.

David Hartmann, Lena Pohlmann, Lelia Hanslik, Noah Gießing, Bettina Berendt, and Pieter Delobelle. Audit me if you can: Query-efficient active fairness auditing of black-box LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33673–33698, 2026. https://aclanthology.org/2026.findings-acl.1681/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. https://arxiv.org/abs/1512.03385.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49(2):1055–1080, 2021. https://doi.org/10.1214/20-AOS1991.

Disi Ji, Robert L. Logan IV, Padhraic Smyth, and Mark Steyvers. Active Bayesian assessment of black-box classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9), pp. 7935–7944, 2021. https://ojs.aaai.org/index.php/AAAI/article/view/16968.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. https://jmlr.org/papers/v17/kaufman16a.html.

Justin Kay, Grant Van Horn, Subhransu Maji, Daniel Sheldon, and Sara Beery. Consensus-driven active model selection. In *IEEE/CVF International Conference on Computer Vision*, 2025. https://arxiv.org/abs/2507.23771.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023. https://arxiv.org/abs/2204.02937.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, PMLR 139, pp. 5753–5763, 2021. https://proceedings.mlr.press/v139/kossen21a.html.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. https://openaccess.thecvf.com/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html.

Patrik Okanovic, Andreas Kirsch, Jannes Kasper, Torsten Hoefler, Andreas Krause, and Nezihe Merve Gürel. All models are wrong, some are useful: Model selection with limited labels. In *International Conference on Artificial Intelligence and Statistics*, PMLR 258, pp. 2035–2043, 2025. https://proceedings.mlr.press/v258/okanovic25a.html.

Vihari Piratla, Soumen Chakrabarti, and Sunita Sarawagi. Active assessment of prediction services as accuracy surface over attribute combinations. In *Advances in Neural Information Processing Systems*, 34, 2021. https://proceedings.neurips.cc/paper/2021/hash/c2f32522a84d5e6357e6abac087f1b0b-Abstract.html.

El Mehdi Saad, Gilles Blanchard, and Nicolas Verzelen. Covariance adaptive best arm identification. In *Advances in Neural Information Processing Systems*, 36, 2023. https://arxiv.org/abs/2306.02630.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. https://arxiv.org/abs/1911.08731.

Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active comparison of prediction models. In *Advances in Neural Information Processing Systems*, 25, 2012. https://proceedings.neurips.cc/paper_files/paper/2012/hash/92fb0c6d1758261f10d052e6e2c1123c-Abstract.html.

scikit-learn developers. `load_digits`: Handwritten digit dataset. Software documentation, accessed September 2026. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems*, 33, 2020. https://arxiv.org/abs/2006.04347.

Tom Yan and Chicheng Zhang. Active fairness auditing. In *International Conference on Machine Learning*, 2022. https://arxiv.org/abs/2206.08450.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *International Conference on Machine Learning*, PMLR 202, pp. 39584–39622, 2023. https://proceedings.mlr.press/v202/yang23s.html.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *International Conference on Machine Learning*, PMLR 235, pp. 62993–63010, 2024. https://proceedings.mlr.press/v235/zrnic24a.html.

A FINITE-POPULATION CONFIDENCE SEQUENCES

A finite-population mixture construction (Waudby-Smith and Ramdas, 2020) gives simultaneous confidence intervals for the fixed, unknown label counts.

A.1 A MIXTURE LIKELIHOOD RATIO FOR AN UNKNOWN COUNT

Consider one stratum containing N fixed binary labels with unknown total K^*. Uniform sampling without replacement reveals a random ordering of these labels. For an ordered response sequence $y_{1:n}$ containing k positives, its probability under a candidate total K is

$$P_K(y_{1:n}) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}\binom{n}{k}}, \tag{14}$$

with probability zero when the observed sequence is inconsistent with K. Let $Q = (N+1)^{-1}\sum_{K=0}^{N} P_K$ denote the uniform mixture over all possible totals. Then

$$Q(y_{1:n}) = \frac{1}{(n+1)\binom{n}{k}}. \tag{15}$$

To verify this identity, draw $U \sim \mathrm{Uniform}[0,1]$ and then draw N conditionally independent Bernoulli(U) labels. Their total is uniform on $\{0,\dots,N\}$, and conditional on that total their ordering is uniform, so these labels follow Q. Integrating over U, a given prefix $y_{1:n}$ has probability $\int_0^1 u^k(1-u)^{n-k}\,du = k!(n-k)!/(n+1)!$, which is (15).

For a total consistent with the observations, define

$$E_n(K) = \frac{Q(y_{1:n})}{P_K(y_{1:n})} = \frac{\binom{N}{n}}{(n+1)\binom{K}{k}\binom{N-K}{n-k}}. \tag{16}$$

Set $E_n(K) = +\infty$ for inconsistent totals. For $\alpha \in (0,1)$, retain

$$\begin{aligned} \mathcal{K}_n(\alpha) &= \left\{K \in \{k,\dots,N-n+k\} : E_n(K) \le \alpha^{-1}\right\}, \\ [l_n, u_n] &= \left[\frac{\min \mathcal{K}_n(\alpha)}{N}, \frac{\max \mathcal{K}_n(\alpha)}{N}\right]. \end{aligned} \tag{17}$$

The accepted set is nonempty. Indeed, because Q is an average of the P_K, at least one K satisfies $P_K(y_{1:n}) \ge Q(y_{1:n})$, and hence $E_n(K) \le 1 < \alpha^{-1}$. Initially $[l_0, u_0] = [0,1]$; at a census it is $[k/N, k/N]$. Each endpoint is an integer count divided by N, and is compatible with all observed labels.

Under the true total, $E_n(K^*)$ is a nonnegative supermartingale starting at one. For a prefix with positive P_{K^*} probability,

$$\mathbb{E}_{K^*}[E_{n+1}(K^*) \mid y_{1:n}] = \frac{\sum_{y:\, P_{K^*}(y_{1:n},y)>0} Q(y_{1:n}, y)}{P_{K^*}(y_{1:n})} \le \frac{Q(y_{1:n})}{P_{K^*}(y_{1:n})} = E_n(K^*). \tag{18}$$

The inequality arises because the mixture may assign mass to continuations that are impossible under the true total. Ville's inequality therefore gives

$$\mathbb{P}_{K^*}\left(\exists n \le N : E_n(K^*) > \alpha^{-1}\right) \le \alpha. \tag{19}$$

Thus K^*/N belongs to every reported interval with probability at least $1-\alpha$, for every fixed total K^*.

A.2 ADAPTIVE INTERLEAVING OF STRATA

The sampling process can be generated by drawing an independent uniform permutation of each stratum before querying starts; selecting a stratum then reveals the next point in its permutation. This construction yields the required sampling law even when the choice of stratum depends on all

previous responses. Hence, on the event that every stratum's sequence covers its true count at every local time, every adaptive interleaving covers all counts at every global time.

For the metadata-based allocation $\alpha_s = \delta p_s / G$,

$$\sum_s \alpha_s = \frac{\delta}{G} \sum_{g=1}^{G} \sum_{s:g(s)=g} p_s = \delta.$$

A union bound over (19) gives

$$\mathbb{P}(\theta^\star \in \mathcal{B}_t \text{ for every global time } t) \geq 1 - \delta, \qquad \mathcal{B}_t = \prod_s [l_s(t), u_s(t)]. \tag{20}$$

All probabilities are over the query randomization, conditional on the fixed labeled pool and its observed metadata.

The intervals can be computed by enumerating feasible totals in (17), using log-binomial coefficients for numerical stability. Inversion examines $N - n + 1$ totals at a new count state; cached states can be reused.

B Exact regret over a binary count box

We prove Theorem 1 for any nonempty product box $\mathcal{B} = \prod_s [l_s, u_s] \subseteq [0,1]^S$.

B.1 Endpoint formulas

The marginal and paired endpoints in the main text are

$$L_{mg} = \sum_{s:g(s)=g} p_s \big[(1 - v_{sm}) l_s + v_{sm}(1 - u_s)\big], \tag{21}$$

$$U_{mg} = \sum_{s:g(s)=g} p_s \big[(1 - v_{sm}) u_s + v_{sm}(1 - l_s)\big], \tag{22}$$

$$B_{mjg} = \sum_{s:g(s)=g} p_s \big[(1 - v_{sm}) v_{sj}(2u_s - 1) + v_{sm}(1 - v_{sj})(1 - 2l_s)\big]. \tag{23}$$

Agreement coordinates contribute zero to B_{mjg}. Its value can be negative, and truncating it at zero would loosen the certificate. Set $\underline{W}_j = \max_h L_{jh}$.

Proof of Theorem 1. Fix two different candidates m, j. Since the group set is finite,

$$\sup_{\theta \in \mathcal{B}} (W_m - W_j) = \max_g \sup_{\theta \in \mathcal{B}} (R_{mg} - W_j). \tag{24}$$

For every feasible θ, both $W_j \geq R_{jg}$ and $W_j \geq \underline{W}_j$. Consequently,

$$R_{mg}(\theta) - W_j(\theta) \leq \min\{B_{mjg},\ U_{mg} - \underline{W}_j\}. \tag{25}$$

To prove attainability, fix a target group g. Within that group choose

$$\theta_s^{(g)} = \begin{cases} u_s, & v_{sm} = 0, \\ l_s, & v_{sm} = 1. \end{cases} \tag{26}$$

This choice maximizes R_{mg}. It also maximizes $R_{mg} - R_{jg}$: coordinates with $v_{sm} = v_{sj}$ have zero difference coefficient, whereas the two disagreement orientations $(v_{sm}, v_{sj}) = (0,1)$ and $(1,0)$ have coefficients $2p_s$ and $-2p_s$, respectively, with the same maximizing endpoints as in R_{mg}. Hence at this corner,

$$R_{mg} = U_{mg}, \qquad R_{mg} - R_{jg} = B_{mjg}, \qquad R_{jg} = U_{mg} - B_{mjg}.$$

In every other group $h \neq g$, independently choose endpoints minimizing candidate j's risk. This is possible because distinct groups have disjoint coordinates. Write $H_{-g} = \max_{h \neq g} L_{jh}$, with $H_{-g} = -\infty$ when $G = 1$. The resulting compatible point satisfies

$$R_{mg} - W_j = U_{mg} - \max\{U_{mg} - B_{mjg}, H_{-g}\} \tag{27}$$

$$= \min\{B_{mjg}, U_{mg} - H_{-g}\}. \tag{28}$$

Moreover, $U_{mg} - B_{mjg} \geq L_{jg}$, since the left side is a feasible value of R_{jg}. Thus $B_{mjg} \leq U_{mg} - L_{jg}$, so replacing H_{-g} by $\underline{W}_j = \max\{H_{-g}, L_{jg}\}$ leaves this minimum unchanged. The bound in (25) is therefore attained. Maximizing over g in (24) yields

$$\sup_{\theta \in \mathcal{B}}(W_m - W_j) = \max_g \min\{B_{mjg}, U_{mg} - \underline{W}_j\}.$$

For $j = m$, the pairwise supremum is zero. Finally,

$$\sup_{\theta \in \mathcal{B}} \mathrm{Reg}_m(\theta) = \sup_{\theta \in \mathcal{B}} \max_j (W_m - W_j) = \max_j \sup_{\theta \in \mathcal{B}} (W_m - W_j),$$

which proves (6).

On event (20), the true parameter belongs to every box, giving $\mathrm{Reg}_m(\theta^\star) \leq C_m(t)$ for every candidate and time. Issuing a certificate with $C_m(t) \leq \tau$ when the actual regret exceeds τ implies failure of that event, whose probability is at most δ. □

For the count intervals in Appendix A, every attaining endpoint is the smallest or largest member of an accepted integer-count set, divided by the stratum size. The attaining corner therefore lies in the product of normalized acceptance sets $\prod_s \{K/N_s : K \in \mathcal{K}_s\}$ and is realized by some labeling of the unqueried points. Because that discrete product is contained in the interval box and contains an attaining corner, the supremum over either set is the same. Taking interval hulls thus causes no optimization gap, even if an acceptance set is not a run of consecutive integers.

The selection guarantee allows arbitrary data-dependent choice among currently certified candidates and any acquisition policy obeying the sampling protocol. The error probability is unconditional over the complete querying and stopping procedure.

B.2 Agreement completion after a disagreement census

Proof of Proposition 1. Once all disagreement labels are known, candidate m's group risk has the form $d_{mg} + c_g$. Each agreement point supplies one shared binary error indicator. Every count between zero and $|A_g|$ can be realized by choosing which common predictions are wrong. Thus the feasible set is exactly

$$\mathcal{C}_A = \prod_{g=1}^{G} \{0, 1/N_g, \ldots, |A_g|/N_g\}, \qquad a_g = |A_g|/N_g.$$

Fix m, j, g, and let $\bar{d}_j = \max_h d_{jh}$. Every completion satisfies

$$d_{mg} + c_g - \max_h (d_{jh} + c_h) \leq \min\{d_{mg} - d_{jg},\ d_{mg} + a_g - \bar{d}_j\}. \tag{29}$$

The first bound uses the competitor's risk in the same group; the second uses $c_g \leq a_g$ and $c_h \geq 0$.

Choose the endpoint completion $c_g = a_g$ and $c_h = 0$ for $h \neq g$. Its objective value is

$$d_{mg} + a_g - \max\{d_{jg} + a_g, \max_{h \neq g} d_{jh}\} = \min\{d_{mg} - d_{jg}, d_{mg} + a_g - \max_{h \neq g} d_{jh}\} \tag{30}$$

$$= \min\{d_{mg} - d_{jg}, d_{mg} + a_g - \bar{d}_j\}. \tag{31}$$

The second equality holds because including d_{jg} in the inner maximum only adds the term $d_{mg} + a_g - d_{jg} \geq d_{mg} - d_{jg}$ to the minimum. When $G = 1$, the maximum over other groups is omitted. Since the completion set and both index sets are finite, their maxima may be interchanged. Therefore

$$\max_{c \in \mathcal{C}_A} \mathrm{Reg}_m(d + c) = \max_{j,g} \min\{d_{mg} - d_{jg}, d_{mg} + a_g - \bar{d}_j\},$$

as asserted. Including $j = m$ makes the maximum nonnegative: for that comparison, a group maximizing d_{mh} gives a zero term. A maximizing pair (j, g) supplies an endpoint completion attaining C_m^D.

A fixed candidate meets the regret tolerance under every completion exactly when its C_m^D is at most τ. Taking the minimum over candidates gives the necessary-and-sufficient condition $\rho_D \le \tau$.

For the randomized assertion, expose all disagreement labels for free. Without agreement queries, a selector's output distribution π is the same for every completion. Define the G endpoint completions $c^{[g]} = a_g e_g$, where e_g is the gth coordinate vector. The attainability argument above shows that every candidate's maximum regret is attained by at least one of these completions. If $\rho_D > \tau$, each candidate therefore fails on at least one endpoint. Writing

$$F_g = \sum_m \pi_m \mathbf{1}\{\mathrm{Reg}_m(d + c^{[g]}) > \tau\},$$

we obtain

$$\sum_{g=1}^{G} F_g = \sum_m \pi_m \sum_{g=1}^{G} \mathbf{1}\{\mathrm{Reg}_m(d + c^{[g]}) > \tau\} \ge \sum_m \pi_m = 1.$$

Hence an endpoint completion has error probability at least $1/G$. Moreover, some candidate m_0 has $\pi_{m_0} \ge 1/M$, and an endpoint on which m_0 fails has error probability at least $1/M$. Combining the two bounds yields $\max_g F_g \ge \max\{1/G, 1/M\} = 1/\min\{M, G\}$. □

Thus, when $\delta < 1/\min\{M, G\}$, a selector that almost surely returns a candidate needs agreement queries to be uniformly δ-correct. The adverse endpoint may depend on the selector's output distribution.

Before a disagreement census, replacing every agreement-coordinate interval by $[0, 1]$ retains simultaneous coverage. At the census, the disagreement coordinates are singletons. The endpoint completions in the proof belong both to the discrete feasible set and to its box relaxation. Thus evaluating Theorem 1 on that box gives exactly C_m^D.

C Oracle complexity separation

C.1 Observation model and minimax definition

Let $\mathcal{P}_G(q)$ contain problems with two fixed binary candidates and G groups whose known group-conditional prediction masses have total disagreement mass at most q in every group. A query chooses a positive-mass group–prediction stratum and receives a fresh independent label from its conditional distribution. Because the masses are known, a draw from a group's mixture can be simulated by first drawing a prediction pattern with its known probability and then querying that stratum.

An admissible algorithm terminates almost surely and returns a candidate with regret at most τ with probability at least $1 - \delta$ under every $P \in \mathcal{P}_G(q)$. The minimax expected query count is

$$T_G^*(q, \tau, \delta) = \inf_{A \text{ admissible}} \sup_{P \in \mathcal{P}_G(q)} \mathbb{E}_P T_A.$$

For each lower-bound pair, all predictions and masses are identical and known to the algorithm. Only the response laws differ.

C.2 The sequential change-of-measure inequality

For two instances P_+, P_- with the same query strata, let $N_s(T)$ count queries to stratum s. For any event E measurable at an almost surely finite stopping time with finite expected query count, the standard adaptive change-of-measure inequality (Kaufmann et al., 2016) is

$$\sum_s \mathbb{E}_+[N_s(T)]\ \mathrm{KL}(P_{+,s}, P_{-,s}) \ge \mathrm{kl}\big(\mathbb{P}_+(E), \mathbb{P}_-(E)\big), \tag{32}$$

where

$$\mathrm{kl}(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1 - x}{1 - y}.$$

Conditioned on the same history, the algorithm's query kernel is identical under both instances and contributes no likelihood-ratio term. Summing the conditional label divergences gives the divergence

of the stopped transcript; data processing to its event indicator gives (32). The stopped statement also follows by truncation. Infinite expected query counts already satisfy the lower bounds below.

If E denotes returning the unique τ-optimal candidate under P_+ and the two instances have disjoint τ-optimal sets, then admissibility implies $\mathbb{P}_+(E) \geq 1-\delta$ and $\mathbb{P}_-(E) \leq \delta$. Monotonicity of binary relative entropy in this region yields the lower bound $\mathrm{kl}(1-\delta, \delta)$ on the right-hand side.

C.3 One group: upper and lower bounds

Let $d \leq q$ be the known disagreement mass. If $d \leq \tau$, either candidate is already τ-optimal. Otherwise, sample from disagreement alone, choosing its orientations with their known conditional masses. Each query reveals a signed loss difference $Z \in \{-1, 1\}$, and the full risk difference is $D = d\mathbb{E}Z$. Selecting according to the sign of $\widehat{D} = d\overline{Z}_n$ can incur regret greater than τ only if $|\overline{Z}_n - \mathbb{E}Z| > \tau/d$. Hoeffding's inequality gives

$$\mathbb{P}\big(|\overline{Z}_n - \mathbb{E}Z| > \tau/d\big) \leq 2\exp\left(-\frac{n\tau^2}{2d^2}\right).$$

Thus

$$n = \left\lceil \frac{2q^2}{\tau^2} \log \frac{2}{\delta} \right\rceil \tag{33}$$

queries suffice uniformly, and all are disagreement queries.

For the matching lower bound, use an agreement stratum of mass $1-q$ with the same label law under both instances and a disagreement stratum of mass q with predictions $(0,1)$. In the latter take label means $1/2+\varepsilon$ and $1/2-\varepsilon$, where $\varepsilon = \tau/q \leq 1/4$. Their full risk differences are $+2\tau$ and -2τ, respectively. Only disagreement queries distinguish them. The elementary Bernoulli divergence inequality gives

$$\mathrm{KL}(\mathrm{Ber}(1/2+\varepsilon), \mathrm{Ber}(1/2-\varepsilon)) \leq \frac{(2\varepsilon)^2}{(1/2-\varepsilon)(1/2+\varepsilon)} \leq \frac{64}{3}\varepsilon^2. \tag{34}$$

Applying (32) yields

$$\mathbb{E}_+ N_D(T) \geq \frac{3q^2}{64\tau^2}\,\mathrm{kl}(1-\delta, \delta). \tag{35}$$

For $\delta \leq 1/10$, this binary divergence is bounded above and below by universal positive multiples of $\log(1/\delta)$. Equations (33)–(35) prove the one-group rate.

C.4 Two groups: information in agreement labels

In each group, assign mass $1-q$ to prediction $(0,0)$. The remaining mass q has predictions $(0,1)$ in group 1 and $(1,0)$ in group 2, with label zero in both cases. Put $a = 1/4$ and $h = 2\tau$. Under P_+ and P_-, respectively, group 1's agreement label has mean

$$p_+ = \frac{a+h}{1-q}, \qquad p_- = \frac{a-h}{1-q}.$$

Group 2's agreement mean is $a/(1-q)$ in both instances. The assumptions $q \leq 1/4$ and $\tau \leq q/4$ imply $h \leq q/2$ and make all these means strictly between zero and one. With rows indexing candidates and columns indexing groups, the risk matrices are

$$R^{\pm} = \begin{pmatrix} a \pm h & a+q \\ a \pm h + q & a \end{pmatrix}.$$

Under P_+, the worst-group risks are $a+q$ and $a+q+h$; under P_-, they are $a+q$ and $a+q-h$. The unique τ-optimal candidate switches, with gap $h = 2\tau$.

Only group 1 agreement responses have different distributions. Since $p_-(1-p_-) \geq 1/16$ and $p_+ - p_- \leq 8h/3$,

$$\mathrm{KL}\left(\mathrm{Ber}(p_+), \mathrm{Ber}(p_-)\right) \leq \frac{(p_+-p_-)^2}{p_-(1-p_-)} \leq \frac{1024}{9}h^2 \leq 128h^2 = 512\tau^2. \tag{36}$$

Therefore (32) implies

$$\mathbb{E}_+ N_{1,A}(T) \geq \frac{\mathrm{kl}(1-\delta,\delta)}{512\tau^2}. \tag{37}$$

Disagreement-only transcript laws are identical under the two instances, so any policy returning a candidate without agreement queries has error at least $1/2$ on one of them.

C.5 A MATCHING TWO-GROUP UPPER BOUND

Draw n independent labels uniformly within each group and observe both candidates' losses. Hoeffding's inequality and a union bound over the four risk means give simultaneous error at most $\tau/2$ with probability at least $1-\delta$ when

$$n = \left\lceil \frac{2}{\tau^2} \log \frac{8}{\delta} \right\rceil.$$

Each empirical group maximum then differs from its true value by at most $\tau/2$. Selecting the candidate with the smaller empirical maximum has regret at most τ. Its total cost is $2n$. Together with (37), this proves the two-group rate and completes the proof of Theorem 2.

For two candidates, $|R_{1g} - R_{2g}| \leq q$ in every group, and hence $|W_1 - W_2| \leq q$. If $q \leq \tau$, either candidate is acceptable without labels. Duplicating candidates or group distributions embeds the hard pair in larger classes, carrying the same lower bound in q, τ, and δ.

D FINITE-POOL TRANSFER AND FURTHER STRUCTURAL PROPERTIES

D.1 A FINITE-POOL AGREEMENT LOWER BOUND

Proposition 2 (Finite-pool agreement lower bound)**.** *Let $0 < q \leq 1/4$, $0 < \tau \leq q/8$, and $0 < \delta \leq 1/10$. Suppose each of two groups has n points, with qn an integer. There are prediction patterns with disagreement fraction q in each group such that, if*

$$(1-q)n \geq 2\tau^{-2}\log(4/\delta),$$

every selector that terminates and is δ-correct on every labeling of those patterns has some fixed labeled pool on which its expected agreement-query count is at least

$$\frac{\mathrm{kl}(1-2\delta,2\delta)}{2048\tau^2}.$$

Proof. Fix the metadata from the two-group oracle construction, now with n points per group and $A = (1-q)n$ agreement points in each. All disagreement labels are zero. Put $a = 1/4$ and $h = 4\tau$. To obtain a worst-case fixed pool, introduce two auxiliary distributions on its hidden labels. In group 1, agreement labels are independent Bernoulli variables with mean $(a+h)/(1-q)$ under the first distribution and $(a-h)/(1-q)$ under the second. In group 2 they are independent Bernoulli variables with mean $a/(1-q)$ under both distributions. All label coordinates are independent under each auxiliary distribution. The selector's guarantee holds for every fixed realization.

Let $\overline{Y}_g$ denote the empirical agreement mean in group g, and let p_g denote its auxiliary mean. Under either distribution,

$$\mathbb{P}\left(\max_{g=1,2}(1-q)|\overline{Y}_g - p_g| > \tau/2\right) \leq 4\exp\left(-\frac{A\tau^2}{2(1-q)^2}\right) \leq \delta. \tag{38}$$

The last inequality follows from the stated condition on A. On the complementary event, all four finite-pool group risks lie within $\tau/2$ of their auxiliary population values. The same holds for each worst-group risk. Their population gap is 4τ, so their finite-pool gap is at least 3τ, with the same winner. Any τ-optimal output identifies the auxiliary sign on this event.

The selector's error probability, averaged over either auxiliary distribution, remains at most δ. Adding the concentration failure shows that the event E of returning candidate 1 satisfies

$$\mathbb{P}_+(E) \geq 1-2\delta, \qquad \mathbb{P}_-(E) \leq 2\delta.$$

Under an auxiliary distribution, every unqueried label retains its independent Bernoulli law conditional on the adaptive query history. Repeated queries to an already observed label reveal no new information and can be omitted; at most $2n$ distinct points are queried. Only group 1 agreement labels contribute transcript divergence. Consequently,

$$\mathrm{KL}(P_+^{\mathrm{tr}}, P_-^{\mathrm{tr}}) = \mathbb{E}_+[N_{1,A}]\ \mathrm{KL}\left(\mathrm{Ber}(p_+), \mathrm{Ber}(p_-)\right).$$

The assumption $\tau \le q/8$ gives $h \le q/2$, so the bound used in (36) applies and now gives a per-query divergence of at most $128h^2 = 2048\tau^2$. Data processing to E yields

$$\mathbb{E}_+[N_{1,A}] \ge \frac{\mathrm{kl}(1-2\delta, 2\delta)}{2048\tau^2}.$$

This expectation averages the selector's conditional expected query counts over finitely many fixed labelings. At least one such labeling attains or exceeds the average, establishing the claim. □

This transfer applies above the stated size threshold. The completion criterion applies at every pool size.

D.2 Recovery under a common resolved worst group

Proposition 3 (Recovery of disagreement cancellation)**.** *Suppose a nonempty count box has a group g_* satisfying $L_{mg_*} \ge U_{mh}$ for every candidate m and every group $h \ne g_*$. Then*

$$\sup_{\theta\in\mathcal{B}} \mathrm{Reg}_m(\theta) = \max_j B_{mjg_*} \qquad \textit{for every } m.$$

Proof. The assumed inequalities imply $R_{mg_*}(\theta) \ge R_{mh}(\theta)$ for every feasible θ, candidate m, and group h. Hence $W_m(\theta) = R_{mg_*}(\theta)$ simultaneously for all candidates, and

$$\sup_{\theta\in\mathcal{B}} \mathrm{Reg}_m(\theta) = \sup_{\theta\in\mathcal{B}} \max_j (R_{mg_*} - R_{jg_*}) = \max_j B_{mjg_*}.$$

Every agreement coordinate has zero coefficient in these differences. □

The condition requires the same worst group for all candidates throughout the box. Different resolved worst groups generally leave absolute risk levels relevant. Shrinking the box preserves the condition once resolved.

D.3 Paired-plus-level inference for general losses

Suppose a simultaneous event provides $L_{mg} \le R_{mg} \le U_{mg}$ and $R_{mg} - R_{jg} \le B_{mjg}$ for all indices. For a group g attaining W_m,

$$W_m - W_j = R_{mg} - \max_h R_{jh} \le \min\{B_{mjg},\ U_{mg} - \max_h L_{jh}\}.$$

Maximizing over groups and competitors, and including the zero self-comparison, proves validity of the certificate formula from those component bounds alone. This argument also applies to multiclass and bounded losses. Binary product-box geometry supplies the stronger conclusion of simultaneous attainability.

Using the same L, U, B, paired-plus-level inference and Theorem 1 give the same certificate. The paired bounds retain the shared-loss dependence discarded by marginal-only bounds.

A multiclass counterexample. Consider labels in $\{1, 2, 3\}$ and two candidates. In group 1, candidates 1 and 2 predict labels 1 and 2, respectively, and the class-probability region is

$$\theta(t) = \left(\frac{t}{5}, \frac{1}{10} + \frac{7t}{10}, \frac{9}{10} - \frac{9t}{10}\right), \qquad 0 \le t \le 1.$$

Thus $R_{11} = 1 - t/5$ and $R_{21} = 9/10 - 7t/10$. In group 2 take fixed risks $R_{12} = 0$ and $R_{22} = 1/2$, which are realizable by two equal-sized deterministically labeled prediction strata. The component

bounds are $U_{11} = 1$, $B_{121} = 3/5$, and $\underline{W}_2 = 1/2$. The paired-plus-level formula returns $1/2$ for candidate 1, whereas its exact regret supremum is

$$\max_{0\le t\le 1} \min\left\{\frac{1}{10} + \frac{t}{2}, \frac{1}{2} - \frac{t}{5}\right\} = \frac{27}{70} < \frac{1}{2},$$

attained at $t = 4/7$. The level maximum and difference maximum require incompatible values of t. A 70-point first group realizes the same gap on the discrete grid $t = k/7$ with class counts $(2k, 7+7k, 63-9k)$ for $k = 0, \dots, 7$. The correlated class-probability region prevents simultaneous attainment of the component bounds.

D.4 Error allocation and running intersections

Any allocation fixed from observed metadata with $\alpha_s \ge 0$ and $\sum_s \alpha_s \le \delta$ supports the coverage union bound, provided coordinates assigned zero error are left unrestricted. For one group, write D for the nonconstant prediction strata and $q_D = \sum_{s\in D} p_s$. If $q_D > 0$, assigning $\alpha_s = \delta p_s / q_D$ on D and keeping agreement coordinates in $[0, 1]$ is valid for regret, since only disagreement enters each risk difference. If $q_D = 0$, all candidates have identical risks. With multiple groups, the completion criterion determines whether agreement labels can be omitted after the disagreement labels are known.

For any time-uniform interval sequence, define

$$\widetilde{l}_s(t) = \max_{u\le t} l_s(u), \qquad \widetilde{u}_s(t) = \min_{u\le t} u_s(u).$$

On the original simultaneous coverage event, these intersections contain the true coordinates at every time. Whenever the intersected box is nonempty, its exact regret supremum is no larger than that of the current box. On the simultaneous coverage event, intersections therefore preserve or advance first certification along a fixed query path. An empty intersection triggers abstention. Using the intersections to guide acquisition may change the query path.

E From an independently sampled audit pool to population regret

Proposition 4 (Sampling lift)**.** *Suppose audit points are independent draws within each of G predefined groups, with $N_g \ge 1$ points in group g. The M candidates are fixed independently of the audit sample. For $\beta \in (0, 1)$, define*

$$\epsilon_g = \sqrt{\frac{\log(2MG/\beta)}{2N_g}}, \qquad \epsilon = \max_g \epsilon_g.$$

With probability at least $1 - \delta - \beta$, every output certified by Theorem 1 at tolerance τ has population worst-group regret at most $\tau + 2\epsilon$.

Proof. Hoeffding's inequality and a union bound give $|R^{\mathrm{pool}}_{mg} - R^{\mathrm{pop}}_{mg}| \le \epsilon_g$ simultaneously for all MG risks with probability at least $1 - \beta$. Both taking the maximum over groups and taking the minimum over candidates are 1-Lipschitz in the sup norm. Hence on this event,

$$\mathrm{Reg}^{\mathrm{pop}}_m \le \mathrm{Reg}^{\mathrm{pool}}_m + 2\epsilon \qquad \text{for every } m.$$

Conditional on every realized labeled pool, the finite-pool probability of an incorrect certificate is at most δ. Integrating over the pool draw and combining the two failure events proves the assertion, including adaptive acquisition, selection, and stopping. □

The lift assigns failure probability δ to label queries and β to audit sampling. Its sampling assumptions and smallest group size determine the additional population error 2ϵ. Changes in group proportions preserve conditional group risks; changes within groups require assumptions on the shifted distributions.

F ACQUISITION POLICIES AND IMPLEMENTATION

The seven in-house policies receive the same group–prediction metadata and requested-label access. They use the allocation $\alpha_s = \delta p_s/G$, the same finite-count intervals, the certificate in Theorem 1, and the same first-certification stopping rule. Acquisition chooses a stratum; the simulator reveals a uniformly selected remaining point within that stratum. Thus acquisition changes efficiency without changing the certificate's validity.

F.1 WORKING POSTERIOR AND SHARED DECISIONS

The acquisition heuristics place independent uniform working priors on the total positive counts $K_s \in \{0, \dots, N_s\}$. After observing (n_s, k_s), their next-label predictive probability and posterior mean finite-population proportion are

$$\eta_s = \frac{k_s + 1}{n_s + 2}, \qquad \bar{\theta}_s = \frac{k_s + (N_s - n_s)(k_s + 1)/(n_s + 2)}{N_s}. \tag{39}$$

These are working-model quantities; the frequentist guarantee treats every audit label as fixed. At certification, choose among candidates satisfying $C_m \le \tau$ by the smallest $\max_g R_{mg}(\bar{\theta})$, then by model index. Before certification, the primary experiments provisionally recommend the smallest C_m, breaking ties by the same estimated risk and then model index. Comparisons use numerical tolerance 10^{-12}; retrospective violations use 10^{-9} to exclude floating-point noise.

F.2 SEVEN MATCHED ACQUISITION POLICIES

Uniform. Choose stratum s with probability $(N_s - n_s)/\sum_r (N_r - n_r)$. Combined with uniform sampling inside the selected stratum, this is uniform sampling without replacement from the remaining pool.

Disagreement. Apply Uniform to the remaining points whose candidate prediction vector is nonconstant. If none remain and no candidate is certified, abstain. Theorem 2 and Proposition 1 characterize the resulting information obstruction.

CAWS (Certified Agreement-Aware Worst-Group Selection). Sample uniformly from the remaining disagreement points until they are exhausted, then uniformly from the remaining pool. Stop at the first certificate. Every remaining point is reachable, so a census guarantees termination. Because CAWS uses its own acquisition random-number stream (Appendix G), its finite-repeat averages can differ from those of Disagreement even when the fallback is never used.

Mix10. While disagreement points remain, choose Uniform over the entire remaining pool with probability 0.10 and uniform disagreement sampling with probability 0.90. After disagreement exhaustion, use Uniform. The mixing weight was fixed before the experiments.

Variance. Under the working count priors, query the stratum with the greatest expected reduction in the sum of posterior marginal group-risk variances. The common factor M can be removed, giving score

$$p_s^2 \eta_s (1 - \eta_s) \frac{(N_s + 2)^2}{N_s^2 (n_s + 3)^2}.$$

The two possible next posterior means of θ_s differ by $(N_s + 2)/[N_s(n_s + 3)]$. The law of total variance gives the displayed expected variance reduction. This score targets marginal group-risk precision.

Bayes-IG. For each of 128 posterior samples, independently draw

$$U_s \sim \mathrm{Beta}(k_s + 1, n_s - k_s + 1), \qquad K_s = k_s + \mathrm{Binomial}(N_s - n_s, U_s).$$

Compute the worst-group optimal candidate for the sampled count vector, breaking ties by model index. For each available stratum, its next-label probability in that draw is $(K_s - k_s)/(N_s - n_s)$. Estimate the mutual information between the next label and the optimal candidate's identity; query the maximizing stratum, breaking ties by stratum index. Logarithms are natural and a fresh posterior batch is sampled at each step using the method-specific generator.

WAVE. Worst-group Active Validation with Exact certificates minimizes the expected next value of $V_t = \min_m C_m(t)$. If $V_t^{s,y}$ is the value after a hypothetical response y in stratum s, choose

$$s_{t+1} \in \operatorname*{arg\,min}_{s:n_s<N_s} \left[(1-\eta_s)V_t^{s,0} + \eta_s V_t^{s,1}\right]. \tag{40}$$

Scores within 10^{-12} of the minimum are ties; resolve them by the largest $p_s(u_s - l_s)/(n_s + 1)$, then stratum index. One-step gains can vanish on discrete or max–min plateaus. Both responses remain possible under the working prior until a stratum is exhausted, and each hypothetical interval is recomputed exactly as after an actual observation.

The implementation orders strata by sorted group values and then lexicographically by prediction vector. Each round evaluates the certificate, stops if eligible, otherwise evaluates both outcomes in every unexhausted stratum, applies (40) and its tie rule, and requests one uniformly selected remaining point. The label-free `Problem` contains metadata; `State` contains accumulated responses.

F.3 Computational cost and access boundary

Computing marginal and pairwise endpoints takes $O(SM^2)$ operations, followed by $O(GM^2)$ for all certificates. A hypothetical response changes one group's contributions. The batched implementation evaluates outcomes of all available actions using $O(SGM^2)$ arithmetic and storage; streaming would trade vectorization for memory. Interval inversion additionally enumerates $N_s - n_s + 1$ feasible totals for each new $(N_s, n_s, k_s, \alpha_s)$ state using log-binomial coefficients, with count-state caching.

The simulation driver owns the complete labels and pre-drawn within-stratum permutations. Acquisition receives only metadata and queried responses. Full labels are used exclusively by the evaluation driver for oracle risks and retrospective diagnostics. The acquisition interface enforces this separation.

G Primary experimental protocol

G.1 Five fixed synthetic populations

Each synthetic setting fixes a complete labeled population and varies query randomization. Its standard errors therefore describe acquisition randomness conditional on that population.

Rare disagreement. $N = 1{,}000$, $M = 2$, $G = 1$: 960 points have predictions $(0, 0)$ and 40 have $(0, 1)$, with respectively 240 and 30 positive labels. The tolerance is 0.01 and the realized optimality gap is 0.02.

One group. $N = 800$, $M = 2$, $G = 1$: 640 agreement predictions $(0, 0)$ and 160 disagreement predictions $(0, 1)$, with respectively 200 and 120 positive labels. The tolerance is 0.04 and the gap is 0.10.

Level trap $+$. Two groups of 400 points and two candidates. Each group contains 320 agreement predictions $(0, 0)$ and 80 disagreement predictions: $(0, 1)$ in group 1 and $(1, 0)$ in group 2. All disagreement labels are zero. The agreement positive counts are 140 and 100. The tolerance is 0.04 and the gap is 0.10.

Level trap $-$. The same metadata and disagreement labels, with group 1's agreement positive count changed to 60. The optimal candidate switches; the gap and tolerance are unchanged.

Diverse. $N = 600$, $M = 6$, with two groups of 300. NumPy's default generator at seed 32017 draws all candidate predictions independently and uniformly from $\{0, 1\}$, then draws labels the same way. This fixed weak-structure control uses tolerance 0.04.

G.2 Digit fitting and observable groups

The digit pools use scikit-learn's `load_digits` dataset (scikit-learn developers, 2026): 1,797 images with 64 features, normalized by 16. The target indicates an original digit of at least five. For each seed in $\{11, 29, 47, 53, 71\}$, a target-stratified split assigns 1,078 examples to training and 719 to auditing.

Table 6: Synthetic labels to certification: means ± standard errors over 40 query randomizations. "Abstain" denotes 40 abstentions. All policies use the exact certificate.

Policy	Rare	One group	Level +	Level −	Diverse
Uniform	333.4 ± 19.0	103.6 ± 7.4	473.9 ± 11.1	455.1 ± 8.8	557.7 ± 1.5
Variance (Hara et al., 2024)	372.2 ± 7.9	111.5 ± 5.8	507.6 ± 9.5	479.6 ± 7.2	562.5 ± 0.7
Bayes-IG (Ji et al., 2021)	14.5 ± 0.5	21.4 ± 1.4	465.2 ± 12.7	457.3 ± 9.6	550.4 ± 1.9
Disagreement (Sawade et al., 2012)	13.8 ± 0.5	20.6 ± 1.3	abstain	abstain	540.2 ± 1.6
Mix10	15.4 ± 0.6	22.8 ± 1.5	466.3 ± 12.5	463.5 ± 10.1	548.8 ± 1.6
WAVE	13.8 ± 0.5	20.6 ± 1.3	474.9 ± 11.4	469.5 ± 13.0	533.8 ± 1.9
CAWS (ours)	13.8 ± 0.5	20.6 ± 1.3	472.8 ± 11.7	465.5 ± 10.6	543.9 ± 1.7

The training nuisance bit equals the target with probability 0.95 independently per point. Within each audit target class, indices are randomly permuted and the first half receives nuisance value one, the remainder zero. The target is used to generate this benchmark distribution; the selector observes only the resulting nuisance attribute. The two audit groups have sizes 360 and 359. Each prediction uses the image features concatenated with $\text{scale} \cdot (2\,\text{nuisance} - 1)$.

Logistic regression uses `lbfgs`, at most 2,000 iterations, and tolerance 10^{-9}. Broad uses the 12 combinations of scale in $\{0, 0.5, 2, 8\}$ and $C \in \{0.1, 1, 10\}$. Local fixes scale 2 and uses 12 regularization values from `logspace(-0.3,0.3,12)`. Saved prediction arrays specify labels, train/audit indices, nuisance bits, coefficients, intercepts, and hyperparameters. All query policies operate on these frozen predictions. Numerical refitting may vary with software versions.

Group risks are recomputed from the observable nuisance field `audit_spurious`; the target–nuisance group codes and four-group risks also stored in some arrays are not used.

Table 7: Digit pool properties under the two observable groups. Risks are full-pool proportions; range is the largest minus smallest candidate worst-group risk.

Pool	Seed	Strata	Best worst-group risk	Risk range
Broad	11	24	0.095	0.407
Broad	29	30	0.119	0.382
Broad	47	27	0.103	0.399
Broad	53	26	0.111	0.390
Broad	71	28	0.117	0.385
Local	11	26	0.379	0.075
Local	29	26	0.328	0.079
Local	47	26	0.339	0.065
Local	53	26	0.356	0.086
Local	71	24	0.387	0.050

G.3 Randomization, estimands, and reporting

The first eight hexadecimal digits of SHA-256 applied to `source/case/seed` define a seed key. The within-stratum permutation seed is the key plus the query-replicate index. The acquisition seed is the key plus 100,000, plus the replicate index, plus 1,000,000 times the method index in the ordering Uniform, Variance, Bayes-IG, Disagreement, CAWS, Mix10, WAVE. The runner stores these choices in `experiment_config.json`.

Primary runs use $\delta = 0.05$, with $\tau = 0.01$ for Rare, 0.04 for other synthetic cases, and 0.05 for digits. Budget checkpoints are $0, 16, 32, 64, 128, 256, 512, N$. After certification or abstention, later checkpoints retain the final state; a checkpoint budget can therefore exceed the number of labels actually queried. An abstaining policy's retained provisional model remains uncertified. Certification rates count abstentions as uncertified. Table 6 reports completed-run label costs and marks its two all-abstaining cells as abstentions. Table 3 reports RMTC instead, charging those runs their 800-label cap. CelebA likewise uses RMTC over all attempted runs, charging its full budget to each run ending without a certificate.

Synthetic standard errors are the sample standard deviation divided by $\sqrt{40}$. Digit results first average 25 query repetitions within each fitted pool, then weight the five training seeds equally; the standard

error is the standard deviation of the five means divided by $\sqrt{5}$. Thus fitted pools determine the reporting units, while query repetitions measure acquisition randomness within each pool.

Table 8: WAVE digit costs by training seed, preserving fitted-pool variation.

Pool	Seed	Mean labels	Median labels	Mean agreement labels
Broad	11	173.04	170.0	20.76
Broad	29	145.56	143.0	19.52
Broad	47	160.36	135.0	25.52
Broad	53	149.84	139.0	14.72
Broad	71	221.16	224.0	42.44
Local	11	1.00	1.0	0.00
Local	29	54.60	55.0	47.80
Local	47	2.00	2.0	0.00
Local	53	12.84	11.0	0.00
Local	71	1.00	1.0	0.00

Execution used four Python workers and one BLAS thread per worker. Mean WAVE trajectory times were approximately 0.25 seconds on Broad, 0.023 on Local, and 2.02 on Diverse; corresponding Bayes-IG times were 0.14, 0.012, and 1.16 seconds. These end-to-end measurements include acquisition and certificate updates in a shared execution environment.

G.4 Complete internal-policy comparisons

Tables 9 and 10 report the complete digit and Adult results, including fixed-budget regret and query-loop time, under the protocols and replication units described above. Table 4 gives the completion diagnostic for all pools, including the ten CelebA banks.

H Independent implementation checks

We check interval coverage, certificate optimization, completion geometry, and issued decisions independently of the analytical proofs.

Finite-population coverage enumeration. For each $N = 1, \ldots, 9$, every total positive count, and $\alpha \in \{0.01, 0.05, 0.20\}$, enumerate all binary orderings with that count and test exclusion of the true count at any time. Across 162 population-composition/error-level combinations, every exact failure probability is at most its assigned α. The largest at $\alpha = 0.05$ is $1/63 \approx 0.016$; the overall maximum $11/126 \approx 0.087$ occurs at $\alpha = 0.20$. Both maxima occur at $(N, K) = (9, 4)$ or $(9, 5)$.

Independent certificate optimization. For 24 random boxes with two groups and three or four candidates, independently maximize $R_{mg} - z$ subject to the box and $z \geq R_{jh}$ for every h, then maximize over g. Across 216 ordered-pair comparisons, this linear-programming calculation differs from the closed form by less than 4×10^{-16}.

Completion diagnostic. The script `agreement_completion.py` computes d from disagreement labels and a from metadata, retaining every candidate's C_m^D, d_{mg}, and a_g. Fields using complete labels are prefixed `oracle_` and serve retrospective difficulty summaries. The diagnostic operates directly on frozen arrays. On 180 small random pools with two to four candidates and one to three groups, exhaustive enumeration covers 4,756 agreement completions. Direct regret maximization, the completion formula, the G attaining completions, and the box implementation agree within 1.12×10^{-16}.

Anytime trajectory audit. At every visited round, the evaluation driver compares all candidate certificates with full-pool regrets. Of 3,150 primary synthetic and digit trajectories across the seven in-house policies, 3,070 certify and 80 abstain. One trajectory has intermediate regret undercoverage, and zero issued certificates are incorrect. The audit therefore tracks both confidence-bound failures along a path and failures of the final certified decision.

Table 9: Digits: certification labels, fraction of the 719-point pool, and regret at budget 128 (all lower is better). Standard errors use five training-seed means with 25 query repeats each. The seven policies share the count intervals and exact certificate.

	Broad			Local		
Policy	Labels	Pool %	Regret@128	Labels	Pool %	Regret@128
Uniform	224.5 ± 15.4	31.2	0.017 ± 0.007	56.2 ± 10.1	7.8	0.034 ± 0.005
Variance (Hara et al., 2024)	206.4 ± 12.7	28.7	0.010 ± 0.006	255.7 ± 13.3	35.6	0.036 ± 0.006
Bayes-IG (Ji et al., 2021)	175.7 ± 7.5	24.4	0.007 ± 0.005	12.3 ± 2.7	1.7	0.025 ± 0.006
Disagreement (Sawade et al., 2012)	116.8 ± 7.3	16.2	0.003 ± 0.001	6.1 ± 1.7	0.9	0.033 ± 0.005
Mix10	122.4 ± 7.8	17.0	0.004 ± 0.001	6.8 ± 2.1	0.9	0.033 ± 0.005
WAVE	170.0 ± 13.6	23.6	0.103 ± 0.046	14.3 ± 10.3	2.0	0.036 ± 0.006
CAWS (ours)	116.1 ± 7.0	16.1	0.004 ± 0.001	5.6 ± 1.8	0.8	0.034 ± 0.005

Table 10: Adult: certification labels, audit fraction, held-out worst-group error, and query-loop seconds (all lower is better). Standard errors use five split means with ten query repeats each. All 350 runs of the seven in-house policies certify correctly.

Policy	Labels	Pool %	Test worst-group error	Seconds
Uniform	686.0 ± 19.4	68.6	0.188 ± 0.001	0.297 ± 0.020
Variance (Hara et al., 2024)	712.4 ± 16.5	71.2	0.187 ± 0.000	0.286 ± 0.029
Bayes-IG (Ji et al., 2021)	135.6 ± 6.6	13.6	0.188 ± 0.001	0.224 ± 0.018
Disagreement (Sawade et al., 2012)	103.9 ± 5.6	10.4	0.188 ± 0.001	0.041 ± 0.003
Mix10	115.3 ± 5.3	11.5	0.187 ± 0.000	0.044 ± 0.003
WAVE	273.5 ± 29.1	27.3	0.187 ± 0.000	2.257 ± 0.272
CAWS (ours)	104.9 ± 5.0	10.5	0.187 ± 0.000	0.040 ± 0.003

Additional independent mathematical checks. A separate diagnostic, `theory_checks.py`, tests the theoretical statements on generated boxes and pools. Across 100 random boxes, 1,026 ordered-pair linear programs agree with the certificate to maximum absolute error 3.33×10^{-16}. Across 300 random finite pools, direct enumeration of 9,256 shared-error count completions agrees with the completion formula, endpoint witnesses, and linear-time evaluation in (8) to error 1.67×10^{-16}. For 116 obstructed pools, including 38 with duplicate candidates, linear programs over randomized output distributions verify the lower bound $1/\min\{M, G\}$. These computations solve the randomized optimization problem separately over the full completion set and its endpoint subset. The same script independently reproduces the 162-case confidence-sequence enumeration above.

I Certificate ablation and mechanism experiments

I.1 Matched-path marginal ablation

The marginal certificate discards dependence between candidate risks while using the same stratum intervals:

$$C_m^{\mathrm{marg}} = \max\left\{0, \max_{j \neq m}\left[\max_g U_{mg} - \max_h L_{jh}\right]\right\}.$$

It is valid on the simultaneous count event and satisfies $C_m \leq C_m^{\mathrm{marg}}$ pointwise. Both certificates see identical Uniform or WAVE query trajectories and record their own first certification. Acquisition continues until both certify, with WAVE using the exact objective throughout. The 300 replay prefixes match their primary-policy stopping times exactly.

This ablation isolates preservation of paired losses within the same uncertainty region and error allocation. A paired-difference-plus-level construction using identical L, U, B is algebraically identical to the exact certificate; an independent componentwise implementation agrees numerically throughout 120 diagnostic replays.

I.2 Agreement-level mechanism sweep

Each fixed population has two groups of 800 points and two candidates. Each group has $800q$ disagreement points, with predictions $(0, 1)$ in group 1 and $(1, 0)$ in group 2, all labeled zero.

Table 11: Matched certificate ablation. Synthetic standard errors use ten query repeats; digit standard errors use five seed means, each based on ten repeats.

	Uniform trajectory		WAVE trajectory	
Population	Exact	Marginal	Exact	Marginal
Rare disagreement	373.6 ± 45.8	931.4 ± 2.7	14.5 ± 1.0	957.8 ± 5.3
One group	105.2 ± 13.1	450.5 ± 10.9	19.2 ± 1.8	609.2 ± 12.4
Level trap +	443.3 ± 27.0	453.8 ± 21.0	461.3 ± 22.6	461.3 ± 22.6
Level trap −	477.2 ± 12.4	477.2 ± 12.4	485.9 ± 30.0	491.4 ± 26.4
Diverse	556.7 ± 3.8	568.9 ± 2.1	532.3 ± 3.4	573.5 ± 1.9
Broad	225.6 ± 17.0	665.3 ± 1.4	174.8 ± 11.1	644.1 ± 4.1
Local	58.5 ± 12.1	670.8 ± 1.7	14.1 ± 10.3	665.3 ± 1.0

Table 12: Mechanism sweep: labels ± standard error over 20 repeats. Agreement share is WAVE's mean per-trajectory agreement fraction. Uniform, CAWS, and WAVE always certify; Disagreement always abstains.

q	h/q	Uniform	**CAWS (ours)**	WAVE	Agreement share
0.025	-0.75	1539.0 ± 4.4	1531.7 ± 5.5	1492.3 ± 6.7	97.3%
0.025	-0.25	1581.5 ± 1.4	1582.0 ± 1.4	1569.2 ± 3.3	97.5%
0.025	0.00	1593.2 ± 0.6	1593.5 ± 0.5	1588.7 ± 1.5	97.5%
0.025	0.25	1584.0 ± 0.8	1583.7 ± 1.0	1568.7 ± 2.8	97.4%
0.025	0.75	1548.4 ± 3.6	1547.8 ± 3.4	1514.5 ± 5.0	97.4%
0.050	-0.75	1395.2 ± 20.3	1366.0 ± 19.4	1330.2 ± 14.9	94.0%
0.050	-0.25	1549.9 ± 2.9	1549.6 ± 2.8	1495.5 ± 9.6	94.6%
0.050	0.00	1585.7 ± 0.6	1585.0 ± 0.7	1561.5 ± 8.4	94.9%
0.050	0.25	1555.2 ± 3.0	1551.0 ± 3.8	1520.1 ± 7.0	94.7%
0.050	0.75	1391.3 ± 17.2	1365.7 ± 18.5	1334.8 ± 18.0	94.0%
0.100	-0.75	1044.3 ± 25.2	1008.0 ± 33.7	998.6 ± 36.1	83.4%
0.100	-0.25	1448.9 ± 8.4	1426.0 ± 14.2	1409.2 ± 12.1	88.6%
0.100	0.00	1567.2 ± 1.6	1566.1 ± 1.5	1524.7 ± 6.5	89.5%
0.100	0.25	1455.4 ± 9.9	1448.5 ± 9.7	1359.5 ± 37.6	88.0%
0.100	0.75	1032.2 ± 28.3	1017.5 ± 26.9	1039.2 ± 24.1	84.4%
0.200	-0.75	463.9 ± 20.9	577.9 ± 18.3	577.5 ± 16.4	43.7%
0.200	-0.25	1138.1 ± 14.1	1153.6 ± 16.7	1149.5 ± 20.3	72.0%
0.200	0.00	1513.9 ± 5.0	1508.3 ± 6.3	1442.2 ± 9.0	77.8%
0.200	0.25	1173.2 ± 17.8	1159.0 ± 30.3	1160.6 ± 30.7	72.0%
0.200	0.75	467.9 ± 30.0	559.4 ± 18.9	577.6 ± 17.2	43.6%

Other predictions are $(0, 0)$. Agreement positive counts are $200 + 800h$ and 200. The grid uses $q \in \{0.025, 0.05, 0.10, 0.20\}$, $h/q \in \{-0.75, -0.25, 0, 0.25, 0.75\}$, and $\tau = q/8$; all counts are integers. Here h is signed and includes $|h| > q/2$, so this mechanism study extends beyond the oracle theorem's parameter restriction.

Each setting has 20 query repeats under Uniform, Disagreement, CAWS, and WAVE: 1,600 trajectories. The permutation key uses the first eight hexadecimal digits of SHA-256 applied to `mechanism/q/level_ratio`; the script fixes the string representation and method offsets. Because $\tau = q/8$, comparisons across q jointly vary disagreement mass and absolute precision.

WAVE's agreement share ranges from 43.6% to 97.5%. Small-q cases can approach a census. At $q = 0.20$, its mean cost increases from 577.6 at $h/q = 0.75$ to 1,442.2 at $h = 0$, where the candidates tie. Even at a realized tie, establishing the small regret tolerance consumes most of the pool.

The sweep yields 1,200 certified trajectories and 400 disagreement-only abstentions, with no incorrect certificates or intermediate regret undercoverage. The 300 matched ablations produce 600 certifier records, with no incorrect issued certificates and one record with intermediate undercoverage. We report these overlapping primary prefixes separately from the primary trajectories.

I.3 COMPUTATIONAL SCALING

The timing probes vary $M \in \{8, 16, 32, 64\}$ and $G \in \{2, 4, 8\}$ at $N = 1{,}024$, with three independently generated prediction matrices per configuration. Each point has a shared latent bit drawn uniformly from $\{0, 1\}$; each candidate's prediction flips this bit independently with probability 0.10, and the label flips it independently with probability 0.20. Groups are balanced by index modulo G. A uniform 256-point prefix is revealed before timing certificate and acquisition evaluation.

Each measurement uses a fresh process and one BLAS thread. Table 13 reports median times over three measurements for one full certificate and one complete two-outcome acquisition evaluation. Peak resident memory includes the runtime, arrays, state, and batched temporaries; stratum counts are means. Measurements come from a shared execution environment.

Table 13: Computational scaling at $N = 1{,}024$ after 256 labels. Times are milliseconds and memory is MiB.

M	G	Strata	Certificate	Acquisition	Peak RSS
8	2	149.7	0.33	6.00	67.4
8	4	226.7	0.58	13.27	68.7
8	8	288.0	0.61	14.23	72.6
16	2	490.3	2.01	33.56	77.9
16	4	589.3	2.09	75.39	92.3
16	8	685.0	2.31	120.69	122.5
32	2	932.7	10.97	200.96	149.9
32	4	962.0	8.73	182.15	222.2
32	8	987.7	9.80	515.61	366.9
64	2	1023.7	39.78	655.03	422.9
64	4	1024.0	38.24	924.03	713.9
64	8	1024.0	39.87	2188.37	1297.1

At $M = 64, G = 8$, every point has a unique group–prediction pattern: one acquisition takes about 2.19 seconds and peak memory is about 1,297 MiB. The $M = 8, G = 2$ case takes approximately 6.0 milliseconds and 67 MiB. These costs motivate replacing the current batched evaluation with streaming action evaluations.

J DIAGNOSTIC REPLAYS AND ALTERNATIVE CONFIDENCE REGIONS

J.1 LOOKAHEAD PLATEAUS AND RUNNING INTERSECTIONS

The diagnostic subset replays three original repetitions for each of five synthetic and ten digit pools under Uniform and WAVE, plus three repetitions for each digit pool under CAWS: 120 trajectories. Original first-certification times match the primary records. Each visited state is checked against an independently assembled paired-plus-level certificate. Running-intersection intervals are maintained on the same labels and their first certificate is recorded separately, without feeding them back into acquisition.

For each of 6,988 pre-stopping WAVE rounds, the predicted gain for action s is

$$V_t - \left[(1 - \eta_s)V_t^{s,0} + \eta_s V_t^{s,1}\right].$$

Table 14 distinguishes all gains being zero, no gain being positive, and multiple minimizing actions, using tolerance 10^{-12}. Current intervals can widen, so predicted gains can be negative. The table first averages fractions within trajectories, then weights digit seeds equally.

All running intersections are nonempty in these replays, and their certificates have zero incorrect decisions and zero intermediate regret undercoverage before stopping. The primary undercoverage event lies outside this subset. Because acquisition uses the current intervals throughout, this comparison measures earlier certification on fixed query paths.

Table 14: WAVE diagnostic subset. Stops use current or running-intersection (nested) boxes. Final columns are mean percentages of pre-stopping rounds. Synthetic cells use three repeats; digit cells use three per training seed.

Population	Current stop	Nested stop	All zero %	No gain %	Tied %
Rare disagreement	17.0	17.0	0.0	0.0	0.0
One group	18.7	18.7	0.0	0.0	0.0
Level trap +	442.7	442.3	28.6	39.5	46.7
Level trap −	486.3	479.7	15.9	30.2	32.4
Diverse	529.3	529.3	58.5	58.5	77.8
Broad	153.3	152.1	9.8	19.5	31.4
Local	13.7	13.7	17.9	17.9	56.0

Table 15: Matched digit replays. Early agreement counts agreement queries among the first $\min(T, 50)$ queries; early queries is the mean of this prefix length.

Pool	Policy	Current stop	Nested stop	Early agreement	Early queries
Broad	Uniform	231.7	230.2	23.6	50.0
Broad	WAVE	153.3	152.1	3.5	50.0
Broad	**CAWS (ours)**	117.3	116.8	0.0	50.0
Local	Uniform	62.5	62.5	33.5	37.3
Local	WAVE	13.7	13.7	9.6	12.7
Local	**CAWS (ours)**	6.7	6.7	0.0	6.7

J.2 Group-pooled confidence sequences

The group-pooled control observes a uniformly random ordering of all pool points, which induces sampling without replacement within each group. For point i, let $\ell_{mi} = \mathbf{1}\{f_m(x_i) \neq y_i\}$ and $Z_{mji} = \mathbf{1}\{\ell_{mi} = 1, \ell_{ji} = 0\}$ for $m \neq j$. In each group, maintain finite-count mixture intervals for all M loss indicators and $M(M-1)$ ordered positive-difference indicators, each with error level $\delta/(GM^2)$. A union bound handles dependence among these indicators.

The loss intervals provide L_{mg}, U_{mg}. If L^Z_{mjg}, U^Z_{mjg} bound the mean of Z_{mji}, use

$$B_{mjg} = \min\{d_{mjg},\ U^Z_{mjg} - L^Z_{jmg}\},$$

where d_{mjg} is the metadata-known candidate-pair disagreement fraction. The identity $\ell_m - \ell_j = Z_{mj} - Z_{jm}$ establishes validity. Combining these components with the paired-plus-level formula gives an anytime-valid upper bound. This construction pools across prediction patterns using GM^2 sequences and requires uniform within-group sampling.

Both certifiers receive identical point labels until both have certified. For the synthetic comparison, $N = 1{,}024$, $\tau = 0.05$, and $\delta = 0.05$. The generator matches the timing probe specification, using distinct seeds: a latent fair bit per point, candidate flips with probability 0.1, label flips with probability 0.2, and groups by index modulo G. Each (M, G) setting uses three independently generated pools, with one uniform order each: 36 paired trajectories. These fresh draws account for the stratum-count differences from Table 13. Separately, each of the ten frozen digit pools receives three new uniform orders: 30 paired trajectories. In total there are 66 paths and 132 certifier records.

These 132 records have zero incorrect certificates and zero intermediate undercoverage. The stratum box uses fewer labels on the digit pools and most scaling configurations, while both constructions approach a census in fragmented regimes. The configurations jointly vary candidate sets, group sizes, and the worst-group objective.

A tighter binary pooled control. For binary predictions and zero–one loss, $Z_{mj} + Z_{jm}$ equals the known disagreement indicator pointwise. Its group mean is therefore d_{mjg} exactly, giving the valid refinement

$$B^{\text{tight}}_{mjg} = \min\{d_{mjg},\ 2U^Z_{mjg} - d_{mjg},\ d_{mjg} - 2L^Z_{jmg}\}.$$

With both original ordered intervals retained, this bound is at most $\min\{d_{mjg}, U^Z_{mjg} - L^Z_{jmg}\}$: the original difference bound is the average of the last two terms in the minimum above.

Table 16: Fixed-precision fragmentation study. Singleton percentage counts points alone in their strata. Final columns are mean pool percentages required to certify, with standard errors over three independently generated pools.

M	G	Strata	Singleton %	Stratum box %	Group pooled %
8	2	162.0	7.4	79.8 ± 1.2	83.4 ± 0.7
8	4	213.0	10.6	92.8 ± 1.0	92.3 ± 0.6
8	8	309.7	16.2	92.1 ± 0.6	93.8 ± 0.5
16	2	499.7	36.0	79.1 ± 1.8	81.9 ± 2.1
16	4	581.3	43.6	89.1 ± 0.8	92.4 ± 1.0
16	8	663.3	52.3	92.4 ± 0.9	95.5 ± 0.7
32	2	928.7	85.6	79.4 ± 4.9	88.0 ± 1.5
32	4	957.3	90.0	88.6 ± 0.5	91.6 ± 1.2
32	8	986.0	93.6	92.8 ± 0.8	95.0 ± 0.4
64	2	1024.0	100.0	81.0 ± 0.5	87.8 ± 0.5
64	4	1023.7	99.9	88.7 ± 0.4	93.1 ± 0.5
64	8	1024.0	100.0	93.1 ± 0.5	95.0 ± 0.5

Table 17: Group-pooled digit comparison on new uniform orders. Means $\pm$ standard errors use five seed means, each from three paired orders.

Pool	Stratum box	Group pooled
Broad	219.1 ± 15.1	320.8 ± 60.8
Local	60.2 ± 20.9	283.0 ± 9.1

An alternative construction maintains one interval $[l_{mjg}, u_{mjg}]$ for each $m < j$, intersects it with the known range $[0, d_{mjg}]$, and sets

$$B_{mjg} = 2u_{mjg} - d_{mjg}, \qquad B_{jmg} = d_{mjg} - 2l_{mjg}.$$

Together with the M group-loss intervals, this uses $G[M + \binom{M}{2}]$ sequences. Assigning each sequence error level $\delta/\{G[M + \binom{M}{2}]\}$ gives simultaneous validity under uniform within-group sampling; an empty intersection triggers abstention. The paired-plus-level formula then supplies an anytime regret bound. This alternative changes both the interval sources and their error allocation. Tables 16 and 17 evaluate the original ordered-indicator construction.

J.3 Tolerance and confidence-level sensitivity

CAWS runs on all ten digit pools at nine (τ, δ) combinations, with three paired query repeats each: 270 trajectories. Seeds and point orders are shared across settings within a fitted pool and repeat. Table 18 uses five training-seed means. The discrete Local intervals and small disagreement regions yield identical stopping counts across the three confidence levels at each tolerance. Tightening tolerance substantially increases the label requirement. The grid has zero incorrect certificates and zero intermediate undercoverage.

Table 18: CAWS sensitivity: mean labels $\pm$ standard error across five seed means, each based on three repeats.

Pool	τ	$\delta = 0.01$	$\delta = 0.05$	$\delta = 0.10$
Broad	0.01	247.7 ± 10.0	247.3 ± 10.3	247.3 ± 10.3
Broad	0.02	182.2 ± 7.1	179.1 ± 6.2	177.7 ± 6.5
Broad	0.05	127.2 ± 6.3	114.3 ± 5.7	108.4 ± 4.7
Local	0.01	36.2 ± 2.6	36.2 ± 2.6	36.2 ± 2.6
Local	0.02	28.2 ± 2.7	28.2 ± 2.7	28.2 ± 2.7
Local	0.05	5.1 ± 1.6	5.1 ± 1.6	5.1 ± 1.6

Tolerance specifies acceptable pool regret: $\tau = 0.01$ permits one percentage point. The failure level bounds the probability of ever issuing an incorrect certificate. Both parameters are prespecified for each setting in this grid.

K PROVISIONAL SELECTION AND NATURAL-GROUP EVALUATION

K.1 SAME-TRAJECTORY RECOMMENDATION INTERVENTION

The selector replay uses all 25 original repeats for each of ten digit pools under Uniform, CAWS, Mix10, and WAVE: 1,000 trajectories. At each checkpoint, the minimum-certificate column uses the original recorded recommendation. Two alternative rules minimize $\max_g R_{mg}(\bar{\theta})$ or $\max_g U_{mg}$, with model-index ties. They replace recommendations before certification; at certification, every variant commits to the original certified output and retains it thereafter. All three columns therefore use the same observations, stopping times, and certified outputs.

The replay matches all 8,000 original checkpoint rows in count, primary selection, regret, and certificate to numerical precision. We aggregate these matched-observation interventions by fitted pool. Table 19 shows that posterior-risk selection substantially reduces WAVE's Broad regret at budget 128 using the same observations, while preserving its acquisition cost.

Table 19: Broad regret at budget 128 on identical query paths. Only uncertified recommendations change. Standard errors use five seed means.

Acquisition	Min. certificate	Min. posterior risk	Min. upper risk
Uniform	0.017 ± 0.007	0.005 ± 0.001	0.005 ± 0.001
Mix10	0.004 ± 0.001	0.003 ± 0.001	0.003 ± 0.001
WAVE	0.103 ± 0.046	0.005 ± 0.001	0.006 ± 0.003
CAWS (ours)	0.004 ± 0.001	0.003 ± 0.001	0.004 ± 0.001

For Local, all CAWS, Mix10, and WAVE paths have certified by budget 128, so these recommendation variants coincide. Before the shared stopping time, recommendations from every variant remain provisional.

K.2 LOCAL TOLERANCE AND THE SEED-29 TIE MECHANISM

Table 20: Local geometry before any query, at $\tau = 0.05$. Near-optimal counts and risk ranges are evaluator-only full-label diagnostics.

Seed	V_0	Truly τ-optimal	Risk range
11	0.053	8/12	0.075
29	0.059	8/12	0.079
47	0.059	9/12	0.065
53	0.056	7/12	0.086
71	0.053	11/12	0.050

Every Local pool requires at least one query, with V_0 close to tolerance and most candidates truly eligible. The small label counts certify the requested 0.05 regret threshold. At budget 128, retained-choice mean regret is 0.036 for WAVE and 0.034 for CAWS. Table 18 reports tighter tolerances.

Across all 25 Local seed-29 WAVE paths, 1,195 of 1,365 query rounds are agreement queries: 47.8 per repetition, with range 46–49. Every such query has maximum predicted gain zero within 10^{-12} and multiple minimizing actions. The width fallback chooses all of them. At these rounds, the recorded maximizing candidate–competitor–group component belongs to group 0 and is limited by its paired-difference bound. Nevertheless, 595 agreement queries go to group 0 and 600 to group 1. The fallback thus spends labels on shared levels while the recorded active component is constrained by a paired difference.

K.3 METADATA-ONLY ERROR ALLOCATIONS

The default mass allocation $\delta p_s/G$ is compared with equal allocation δ/S and square-root allocation $\delta\sqrt{p_s}/\sum_r \sqrt{p_r}$. Each is fixed using metadata, sums to δ, and preserves simultaneous validity. Ten digit pools, three allocations, and ten paired repeats give 300 CAWS trajectories, using common point orders within each pool and repeat. All certificates are correct, with zero intermediate undercoverage.

Table 21: CAWS under metadata-only allocations. Errors use five seed means, each from ten paired repeats. These orders differ from the primary experiment.

Allocation	Broad	Local
Mass	118.7 ± 6.0	5.6 ± 1.4
Equal strata	123.7 ± 5.7	5.6 ± 1.4
Sqrt mass	119.1 ± 5.7	5.6 ± 1.4

Mass allocation has the smallest mean label cost among these three choices on Broad; all three have the same mean cost on Local. This comparison uses current intervals, independently of the running-intersection replays.

K.4 Adult fitting, querying, and held-out evaluation

Adult (Becker and Kohavi, 1996) uses the official UCI files, released under CC BY 4.0. The target is income above \$50K; the observed Female/Male attribute defines disjoint groups. Records are unweighted and `fnlwgt` is excluded from the features. The objective is the maximum unweighted error over these two observed groups.

For each seed in $\{11, 29, 47, 53, 71\}$, a label-independent uniform permutation of the 32,561-record `adult.data` file allocates the first 1,000 records to auditing and the remaining 31,561 to fitting. The official 16,281-record `adult.test` file is excluded from fitting, acquisition, recommendation, and stopping. Its labels are used only by the evaluation driver. Splits share the test file and overlapping source records; their standard errors describe variability across these correlated splits.

Four prespecified feature views use all eligible features, remove sex, remove both education fields, or remove capital gain and loss. Each has eight logistic candidates with $C \in \{0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30\}$. Numeric variables are median-imputed and standardized; categorical variables are mode-imputed and one-hot encoded, ignoring unseen categories. Every preprocessing operation is fitted on the fitting partition only. Logistic regression uses `liblinear`, tolerance 10^{-6}, at most 2,000 iterations, and the split seed. Fitting and prediction take 5.3–5.7 seconds per split, with no convergence warnings. The 32 fitted candidates produce 22–23 distinct audit prediction columns. Regularization, grouping, and tolerance were also prespecified.

Table 22: Adult pool geometry, with $N = 1{,}000$, $M = 32$, and $G = 2$. Distinct vectors count audit prediction columns; near-optimal counts are retrospective at $\tau = 0.02$.

Seed	Female/Male	Strata	Distinct vectors	V_0	Risk range	τ-optimal
11	349/651	70	23	0.111	0.034	20/32
29	349/651	78	23	0.101	0.046	16/32
47	344/656	65	22	0.088	0.031	27/32
53	310/690	61	22	0.080	0.036	19/32
71	329/671	74	23	0.107	0.030	19/32

The query experiment fixes $\delta = 0.05$ and $\tau = 0.02$, with ten repeats for each fitted pool and each of seven policies: 350 trajectories and 2,800 checkpoints. Within-stratum orders are shared across policies at each split and repetition. Acquisition receives only predictions, observed groups, and requested responses. Fitting and audit partitions are disjoint. Postprocessing independently recomputes final audit and test risks from frozen predictions.

Test risk is the maximum empirical error over the two test groups. Test regret subtracts the smallest such maximum among the 32 candidates, using an evaluator-only oracle on the held-out test set. Table 23 reports final selection quality.

Full-label diagnostics show that Male is a realized worst group for all 32 candidates in every split, and 16–27 candidates satisfy the true tolerance. Candidate 0 has audit regret between 0 and 0.012 across splits. This post-hoc geometry identifies many acceptable candidates, while certification quantifies the evidence needed to establish an acceptable choice from queried labels.

Table 23: Adult final quality. Test regret uses the held-out candidate oracle. Errors use five split means. Variance, Bayes-IG, and Disagreement use the cited acquisition baselines; CAWS is the proposed policy and WAVE is the one-step certificate-lookahead control.

Policy	Finite-pool regret	Test regret	Certified (%)
Uniform	0.001 ± 0.000	0.002 ± 0.001	100
Variance (Hara et al., 2024)	0.004 ± 0.001	0.001 ± 0.000	100
Bayes-IG (Ji et al., 2021)	0.000 ± 0.000	0.001 ± 0.001	100
Disagreement (Sawade et al., 2012)	0.001 ± 0.000	0.002 ± 0.001	100
Mix10	0.001 ± 0.000	0.001 ± 0.000	100
WAVE	0.002 ± 0.001	0.001 ± 0.000	100
CAWS (ours)	0.001 ± 0.001	0.001 ± 0.000	100

Table 24: Query costs on the ten fixed CelebA banks at $B = 4{,}000$. CR records certificate issuance. RMTC assigns the cap to each uncertified run. Standard errors use ten bank means, each averaging 25 query repeats.

Policy	CR (%)	RMTC
Uniform	100.0	3184.2 ± 64.5
Disagreement (Sawade et al., 2012)	60.0	2351.6
CAWS (ours)	100.0	1385.4 ± 42.1

All 350 Adult trajectories across the seven in-house policies certify correctly, with zero intermediate undercoverage. Uniform uses 68.6% of the audit pool, CAWS 10.5%, and WAVE 27.3%. Disagreement certifies in every run, consistent with the completion diagnostic. Complete query-loop times include repeated certificates, acquisition, updates, and lightweight trace bookkeeping, excluding fitting and initial oracle calculations. The held-out measurements assess selection quality on the test set; a population guarantee additionally requires the audit-sampling assumptions of the population-lift result.

L Fixed visual candidate banks

Task and observation protocol. The CelebA (Liu et al., 2015) target is the binary `Blond_Hair` attribute; the recorded `Male` attribute defines the two observed groups. Each bank contains 16 candidates from the ResNet-50 and ViT-B/16 families, varying standard versus group-balanced training, regularization strength, and checkpoint. Predictions are frozen before label acquisition on 4,000 held-out audit images. Ten separately fitted banks each have 25 query repeats, with $\tau = 0.02$, $\delta = 0.05$, and label cap $B = 4{,}000$. Acquisition observes group memberships, the fixed binary prediction matrix, and requested labels. The three policies use the same count intervals, metadata-based error allocation, and exact certificate.

Completion and querying. Four banks have $\rho_D \in [0.031, 0.044] > 0.02$. Disagreement abstains on these banks and certifies on the other six. CAWS queries no agreement labels on the six unobstructed banks. Across the four obstructed banks, its agreement-query mean and SE are 711.5 ± 36.2; across all ten banks they are 284.6 ± 116.9. The latter summary includes the six zero-valued bank means. These diagnostics describe the ten fixed banks; the query repetitions vary acquisition randomization within a bank.

Disagreement's mean cost on the six banks where it certifies is 1,252.7 labels. Equal repeat counts across banks give its all-run RMTC

$$0.6(1252.7) + 0.4(4000) = 2351.6$$

after rounding. A successful-run standard error is not assigned to this all-run quantity. CAWS has 56.5% lower RMTC than Uniform on these banks. The shared certificate guarantees termination after a completed census; the disagreement-only policy can instead stop without a certificate when its query set is exhausted.

M RELATION TO SEQUENTIAL SELECTION MODELS

Model Selector (Okanovic et al., 2025), the variance-minimization approach (VMA) of Hara et al. (2024), and CODA (Kay et al., 2025) study informative model comparisons with limited labels. The present results characterize which labels can be omitted under a worst-group objective and how shared binary losses constrain regret over a confidence region. These structural questions are distinct from optimizing acquisition for an average-risk comparison. Active fairness auditing (Yan and Zhang, 2022) selects queries to estimate fairness properties under its own target functional and observation protocol.

Maximin action identification (Garivier et al., 2016) studies a related maximum/minimum geometry while sampling selected action-pair rewards. Here a query reveals all candidate binary losses at a known group–prediction pattern. Agreement labels can therefore carry no information about within-group differences and yet determine the worst-group ranking. The oracle theorem isolates where this information resides, without claiming instance-optimal maximin identification.

Covariance-adaptive best-arm identification (Saad et al., 2023) exploits dependence among simultaneously observed arms for mean-reward selection. The present target is a maximum of group means, with finite-pool sampling without replacement in adaptive strata and a binary count-box regret calculation. Distributionally robust selection (Fan et al., 2020) instead compares simulated alternatives under a finite ambiguity set of input distributions. These connections delimit the observation model and target; robust objectives and shared observations themselves are established ideas.