STABLE ACCURACY, UNSTABLE ALLOCATIONS: RETRAINING REPRODUCIBILITY IN SELECTIVE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retraining a selective predictor can change which cases receive review even when accuracy changes little. We study the reproducibility of these review assignments under exact fixed-capacity top-k selection, measuring allocation reproducibility (AR) by mean pairwise normalized review-set overlap. Across six visual settings with ten retrainings each, accuracy standard deviations are 0.11–0.50 percentage points, yet 37.6–50.1% of review slots are reassigned to different cases at a 10% budget. Civil Comments/DistilBERT shows lower turnover, at 11.3%. Changes affect persistent errors and also occur when predicted classes remain unchanged. We derive a sharp bound linking Spearman correlation to top-k replacements, specialize an established subset-stability identity to connect set-level and individual disagreement, and characterize how realized score and cutoff changes constrain membership changes. Two-model probability averaging reduces median allocation turnover across all tested settings and budgets, but does not eliminate it. When review identity matters, accuracy and selective risk alone do not characterize the reproducibility of the resulting allocation.

1 INTRODUCTION

Selective prediction uses uncertainty to decide which cases to defer, review, or abstain on. Existing work primarily evaluates risk–coverage trade-offs (Chow, 1970; El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017; 2019) and predictive reliability (Hendrycks and Gimpel, 2017; Lakshminarayanan et al., 2017; Guo et al., 2017; Ovadia et al., 2019). These quantities describe selective quality, but not whether retraining assigns review to the *same cases*.

This distinction matters under fixed capacity. When only k cases can be reviewed, selecting one case necessarily excludes another. A system therefore has both aggregate outcomes, such as coverage and selective risk, and an *allocation*: the identities of the cases receiving review. Stable aggregate performance does not imply a stable allocation. We ask whether independent retraining preserves this allocation on a fixed evaluation population.

Training randomness makes this question nontrivial. Initialization, minibatch order, augmentation, and implementation nondeterminism can produce different models with similar aggregate performance (Henderson et al., 2018; Summers and Dinneen, 2021; Bouthillier et al., 2021). Prediction churn studies changes in $\hat{y}_{i,s}$ (Bahri and Jiang, 2021), while predictive multiplicity and underspecification examine behavioral differences among similarly performing models (Marx et al., 2020; D'Amour et al., 2022). Prior work also studies prediction consistency across randomized training and uses selective ensembles to reduce prediction disagreement (Black et al., 2022). We instead study the reproducibility of exact fixed-capacity review membership induced by uncertainty rankings. This allocation depends on both sample-level scores and the population-level top-k cutoff, and can change even when the predicted class does not.

Global uncertainty rankings also provide only limited information about exact review membership. A top-k policy is determined by crossings at the operational cutoff, so high global rank agreement can coexist with substantial top-k disagreement. Conversely, large changes elsewhere in the ranking need not affect the selected set. We call this mismatch the *allocation identifiability gap*: aggregate or global summaries may constrain, but need not determine, the induced allocation (Figure 1).

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.

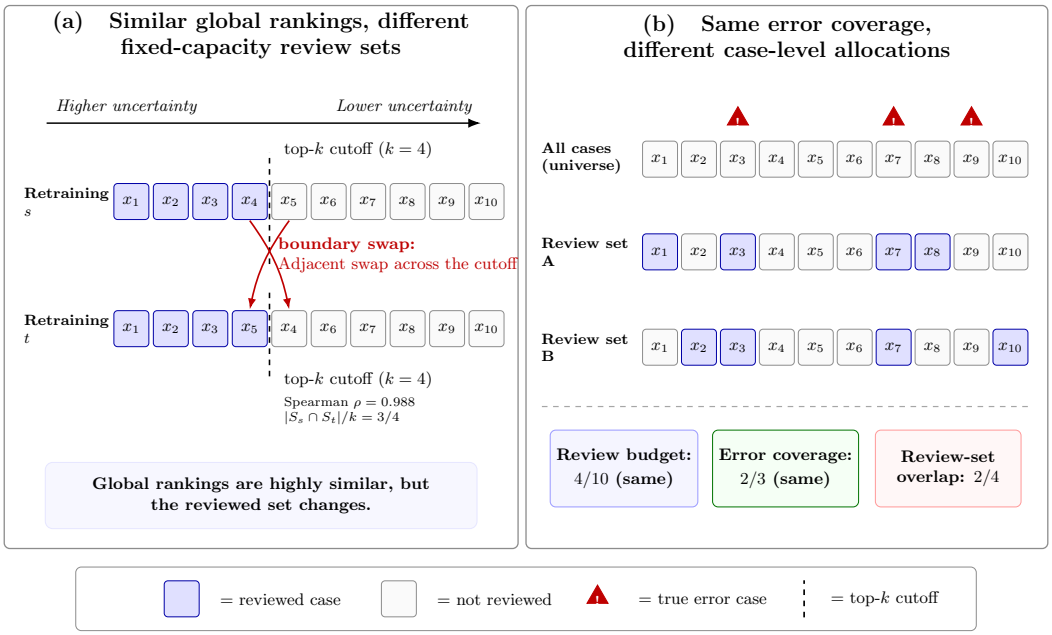

Figure 1: Global ranking agreement and aggregate coverage do not determine fixed-capacity review membership. Similar rankings can yield different review sets, and identical coverage can conceal case-level turnover.

We evaluate seven model–dataset settings: ResNet-18/34 on CIFAR-10/100, DenseNet-121 on CIFAR-10, ResNet-18 on ImageNet-100, and DistilBERT on Civil Comments. At a 10% review budget, mean pairwise turnover is 37.6–50.1% across the six visual settings, despite accuracy standard deviations of only 0.11–0.50 percentage points. Civil Comments shows lower turnover at 11.3%. A same-task R18/C10 control does not reproduce Civil's high AR by sharing initialization alone. Allocation changes affect persistent errors and cases whose predicted classes remain unchanged. Two-model probability averaging reduces median allocation turnover across all tested settings and budgets, but does not eliminate it.

Our contributions are:

- **Allocation reproducibility under retraining.** We quantify fixed-capacity review turnover across seven model–dataset settings, including among persistent errors and cases with unchanged predicted classes.
- **A sharp rank–allocation bound.** We derive a tight bound linking Spearman agreement to exact top-k replacements, characterize the ambiguity left by aggregate coverage, and use an established subset-stability identity (Nogueira et al., 2018) to obtain an exact finite-run decomposition of allocation disagreement into case-level instability.
- **Certificates for allocation stability.** We derive a pair-adaptive, post-hoc certificate that localizes possible membership changes, together with sharp worst-case perturbation bounds, and test how tight these bounds are on observed retraining pairs.

2 Related Work

Selective prediction and uncertainty. Selective classification studies abstention through risk–coverage trade-offs and learned selection (Geifman and El-Yaniv, 2017; 2019). Confidence, calibration, and uncertainty estimation assess predictive reliability (Corbière et al., 2019; Guo et al., 2017; Lakshminarayanan et al., 2017; Ovadia et al., 2019). These evaluations do not by themselves establish whether retraining preserves the cases selected for review at fixed capacity.

Training variability, churn, and multiplicity. Training randomness can produce models with similar aggregate performance but different behavior (Summers and Dinneen, 2021; Bouthillier et al., 2021). Prediction churn measures label disagreement across retrainings (Bahri and Jiang, 2021); predictive multiplicity and underspecification examine differences among similarly performing models (Marx et al., 2020; D'Amour et al., 2022). We study review membership, which can change even when predicted classes remain unchanged.

Subset, ranking, and replicability. Subset stability is well studied in feature selection (Meinshausen and Bühlmann, 2010; Nogueira et al., 2018). AR uses established normalized overlap; for equal-size selections, its chance-corrected form equals the Φ statistic of Nogueira et al. (2018). Fagin et al. (2003) study distances between top-k lists and use related block permutations. We establish the minimum squared full-ranking displacement for a prescribed number of top-k replacements. Recent work studies assumption-free ranking stability (Liang et al., 2025), reproducibility and replicability in learning (Impagliazzo et al., 2022; Hopkins and Moran, 2025), and replicable conformal prediction (Papamichalis et al., 2026). Our setting fixes the evaluation population, varies training randomness, and retains exact selection capacity.

3 Fixed-Capacity Allocation Reproducibility

Let f_{θ_s} denote the classifier from the sth independently randomized execution of a fixed training pipeline, $s = 1, \ldots, S$, with $S \geq 2$. All runs are evaluated on the same n labeled examples (x_i, y_i). Let $\hat{y}_{i,s}$ denote the predicted class and $u_{i,s}$ an uncertainty score, with larger values indicating greater uncertainty. For review budget $\alpha \in (0, 1)$, set $k = \lceil \alpha n \rceil$ and define

$$\mathcal{S}_s^\alpha = \mathrm{TopK}(u_{\cdot,s}, k), \qquad Z_{i,s}^\alpha = \mathbb{I}[i \in \mathcal{S}_s^\alpha]. \tag{1}$$

TopK returns the indices of the k largest scores, breaking ties by ascending sample index. Thus every run selects exactly k cases. We use "review" for any downstream action assigned to selected cases and suppress α when unambiguous.

Allocation reproducibility. We measure set-level agreement by mean normalized overlap:

$$\mathrm{AR}@\alpha = \binom{S}{2}^{-1} \sum_{s<t} \frac{|\mathcal{S}_s^\alpha \cap \mathcal{S}_t^\alpha|}{k}. \tag{2}$$

AR equals one iff all runs select the same cases. The mean number of reassigned review slots is $k(1 - \mathrm{AR})$.

Case-level disagreement. Let $m_i = \sum_s Z_{i,s}$ and $\hat{p}_i = m_i/S$. Define

$$D_i = \frac{m_i(S - m_i)}{\binom{S}{2}}, \qquad \widehat{\mathrm{IDI}}_i = 4\hat{p}_i(1 - \hat{p}_i), \qquad D_i = \frac{S}{2(S-1)} \widehat{\mathrm{IDI}}_i. \tag{3}$$

Here D_i is the fraction of run pairs that disagree on case i. The bounded individual decision instability (IDI) is largest for cases selected in roughly half of the runs. These quantities are rescalings, not independent measures; Section 4.2 relates them to AR.

Common-error diagnostics. For $q \in \{1, \ldots, S\}$, define the common-error universe

$$\mathcal{E}_q = \left\{ i : \sum_s \mathbb{I}[\hat{y}_{i,s} \neq y_i] \geq q \right\}. \tag{4}$$

We use $q = 8$ for the primary ten-run analysis and $q \in \{6, 8, 10\}$ for sensitivity. We report

$$\mathrm{CEFR}@\alpha = \frac{\sum_{i \in \mathcal{E}_q} \mathbb{I}[0 < m_i < S]}{|\mathcal{E}_q|}, \qquad \mathrm{PairDis}(\mathcal{E}_q, \alpha) = \frac{\sum_{i \in \mathcal{E}_q} D_i}{|\mathcal{E}_q|}. \tag{5}$$

CEFR measures the fraction with any observed change in review status; PairDis measures mean pairwise disagreement. Since never-reviewed and always-reviewed cases both contribute zero to

CEFR, we also report the never/flip/always partition (Appendix C.3). CEFR depends on the number of runs and should not be compared across different S without accounting for that dependence.

Common-error coverage and its across-run range are

$$A_s^\alpha = \frac{1}{|\mathcal{E}_q|} \sum_{i \in \mathcal{E}_q} Z_{i,s}^\alpha, \qquad \Delta_{\mathrm{ECR}}^\alpha = \max_s A_s^\alpha - \min_s A_s^\alpha. \tag{6}$$

We call A_s^α common-error coverage (ECR) and report its across-run range $\Delta_{\mathrm{ECR}}^\alpha$. Unlike per-run error coverage, ECR uses the same error population in every run. Common-error metrics are N/A when $\mathcal{E}_q$ is empty.

Prediction agreement. To compare allocation changes with prediction churn, we report

$$\mathrm{PredAgr} = \binom{S}{2}^{-1} \sum_{s<t} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\hat{y}_{i,s} = \hat{y}_{i,t}]. \tag{7}$$

AR and prediction agreement use the full evaluation population; PairDis is conditioned on $\mathcal{E}_q$.

4 Theoretical Characterization

This section formalizes the allocation identifiability gap at three levels: global ranking, case-level disagreement, and realized score perturbations.

4.1 Limits of global summaries

We first ask how much full-ranking agreement can certify about exact top-k membership. Let π_i and σ_i denote the ranks of item i in two strict rankings of N items, with rank 1 selected first. Let S_π^k and S_σ^k be their top-k sets, and define

$$r = k - |S_\pi^k \cap S_\sigma^k|. \tag{8}$$

Theorem 1 (Sharp rank–allocation frontier)**.** *For $N \geq 2$ and $1 \leq k < N$,*

$$\sum_{i=1}^{N} (\pi_i - \sigma_i)^2 \geq 2r^3, \qquad \rho(\pi, \sigma) \leq 1 - \frac{12r^3}{N(N^2-1)}. \tag{9}$$

For every feasible $r \in \{0, \ldots, \min(k, N-k)\}$, equality is attained by exchanging the adjacent length-r blocks immediately on either side of the top-k cutoff while preserving within-block order.

The minimum is therefore attained by exchanging the two blocks adjacent to the cutoff. This construction is related to the block permutations of Fagin et al. (2003); the result establishes its optimality for a prescribed number of top-k replacements.

Corollary 1 (Spearman-certified overlap)**.** *For an observed strict-ranking Spearman correlation ρ, define*

$$r_{\max} = \min\left\{ k, N-k, \left\lfloor \left(\frac{N(N^2-1)(1-\rho)}{12} \right)^{1/3} \right\rfloor \right\}. \tag{10}$$

Then

$$r \leq r_{\max}, \qquad \frac{|S_\pi^k \cap S_\sigma^k|}{k} \geq 1 - \frac{r_{\max}}{k}. \tag{11}$$

The certificate requires the strict permutation induced by the fixed tie-breaking rule; standard average-tie Spearman values should not be substituted directly. When $N \geq 2k$ and $r = k$, disjoint top-k sets can have

$$\rho = 1 - \frac{12k^3}{N(N^2-1)}. \tag{12}$$

For $N = 10{,}000$ and $k = 1{,}000$, this gives $\rho \approx 0.988$ despite zero top-10% overlap.

Aggregate coverage loses identity more directly.

Proposition 1 (Aggregate-allocation non-identifiability). *Let $\mathcal{E}$ contain m items and let $A, B \subseteq \mathcal{E}$ satisfy $|A| = r_1$ and $|B| = r_2$. Then*

$$\max(0, r_1 + r_2 - m) \leq |A \cap B| \leq \min(r_1, r_2), \tag{13}$$

and every integer intersection size in this interval is attainable. In particular, when $1 \leq r_1 = r_2 = r \leq m/2$, identical coverage r/m is compatible with disjoint allocations.

Thus global rank agreement and aggregate coverage can constrain an allocation without identifying its case-level membership.

4.2 From set disagreement to individual cases

Under exact fixed-capacity selection, the subset-stability identity of Nogueira et al. (2018) yields an exact decomposition of allocation disagreement.

Proposition 2 (Exact finite-run allocation decomposition). *If each of $S \geq 2$ runs selects exactly k cases from the same evaluation universe, then*

$$1 - \mathrm{AR} = \frac{1}{2k} \sum_{i=1}^{n} D_i. \tag{14}$$

Case i disagrees on $m_i(S - m_i)$ unordered run pairs. Summing over cases and using the common capacity k gives the identity; no population or independence assumption is required.

Define the finite-S corrected instability

$$\widetilde{\mathrm{IDI}}_i = \frac{S}{S-1} \widehat{\mathrm{IDI}}_i = 2D_i. \tag{15}$$

Then

$$\sum_{i=1}^{n} \widetilde{\mathrm{IDI}}_i = 4k(1 - \mathrm{AR}). \tag{16}$$

Thus aggregate corrected IDI and AR encode the same disagreement, whereas case-level IDI localizes its contribution. Under i.i.d. retraining, $\widetilde{\mathrm{IDI}}_i$ is unbiased for $4p_i(1 - p_i)$. For $0 < k < n$, the chance-corrected quantity

$$\mathrm{AR}_{\mathrm{adj}} = \frac{\mathrm{AR} - k/n}{1 - k/n} \tag{17}$$

equals the subset-stability estimator $\widehat{\Phi}$ of Nogueira et al. (2018). Derivations are given in Appendices A and D.

4.3 Localizing realized membership changes

Let $u, v \in \mathbb{R}^n$ be uncertainty-score vectors on the same evaluation population, and let S_u, S_v be their exact top-k sets under the fixed sample-index tie-breaking rule. Let τ_u and τ_v denote their respective kth-largest scores. Define

$$R_u = \{i : |u_i - \tau_u| \leq |u_i - v_i| + |\tau_u - \tau_v|\}, \tag{18}$$

and define R_v symmetrically.

Proposition 3 (Boundary stability certificate). *For every $1 \leq k \leq n$,*

$$S_u \triangle S_v \subseteq R_u \cap R_v, \qquad 1 - \frac{|S_u \cap S_v|}{k} \leq \min\left\{1, \frac{|R_u \cap R_v|}{2k}\right\}. \tag{19}$$

If $\|u - v\|_\infty \leq \delta$, then $|\tau_u - \tau_v| \leq \delta$, and every membership change lies within 2δ of the corresponding cutoff in each score vector.

The certificate is pair-adaptive and post hoc: it identifies where membership changes can occur for an observed pair of retrainings. Section 5.4 evaluates its tightness on observed pairs, and Appendix A.7 gives the proof.

For comparison, let $u_{(1)} \geq \cdots \geq u_{(n)}$, $1 \leq k < n$, and $1 \leq t \leq \min(k, n-k)$. Define

$$G_t(u) = u_{(k-t+1)} - u_{(k+t)}. \tag{20}$$

Then the sharp worst-case guarantee is

$$\|u - v\|_\infty < \frac{G_t(u)}{2} \quad \Longrightarrow \quad k - |S_u \cap S_v| \leq t - 1. \tag{21}$$

The threshold is sharp: for every $\varepsilon > 0$, there exists a perturbation with norm at most $G_t(u)/2 + \varepsilon$ that causes at least t replacements. Setting $t = 1$ gives the exact-set stability radius. Full statements, tie cases, and proofs are given in Appendix A.5. These bounds describe worst-case perturbations and are not intended to model the variability produced by retraining.

Stable coverage within a fixed subpopulation can coexist with variable individual review decisions. Appendix A gives the variance decomposition and a construction in which negative covariances exactly cancel individual variances.

5 Experiments

5.1 Protocol

Settings. We evaluate ResNet-18/34 on CIFAR-10/100 (Krizhevsky, 2009; He et al., 2016), DenseNet-121 on CIFAR-10 (Huang et al., 2017), ResNet-18 on ImageNet-100 (Deng et al., 2009), and DistilBERT on Civil Comments (Borkan et al., 2019; Sanh et al., 2019), with ten independently randomized runs per setting. Visual models train from scratch: CIFAR uses SGD with cosine decay for 100 epochs (110 for R34/C100), and ImageNet-100 uses 50 epochs. ImageNet-100 contains 50,000 training and 5,000 disjoint evaluation images at 224×224 resolution. DistilBERT is fine-tuned for two epochs from a shared pretrained checkpoint. Full protocols are in Appendix B.

Allocation and statistical analysis. Review sets are formed post hoc from predictive entropy at $\alpha \in \{1, 5, 10, 20, 30\}\%$, using the exact top-$k$ rule and fixed tie breaking defined above. Common-error analyses use $q = 8$, with $q \in \{6, 10\}$ for sensitivity.

AR, prediction agreement, and entropy-rank correlation average all 45 run pairs. Under i.i.d. retraining, these are order-2 U-statistics, so the pairs are not independent replicates. Standard errors are estimated by delete-one-run jackknife, conditional on the fixed evaluation population (Appendix C.1). CEFR and PairDis use the jointly defined $\mathcal{E}_q$ and are therefore excluded from this U-statistic treatment.

5.2 Aggregate Performance and Review-Set Reproducibility

Table 1 summarizes the top-10% results. Across the six visual settings, accuracy standard deviations are 0.110–0.496 percentage points, while AR is 0.4994–0.6238, corresponding to 37.6–50.1% review-slot turnover. This amounts to 376.2–500.6 of 1,000 slots for CIFAR and 227.3 of 500 for ImageNet-100. Civil has lower relative turnover (11.3%), but still reassigns 1,101.3 of 9,732 slots on average. Delete-one-run jackknife standard errors for top-10% AR are 0.0034–0.0051 across the six visual settings and 0.0012 for Civil (Appendix C.1).

The reported entropy Spearman values use average ranks for ties, whereas Theorem 1 applies to strict rankings after deterministic tie breaking. Visual entropy-rank agreement is 0.6377–0.8716 but does not determine exact allocation: ImageNet-100 has $\rho = 0.8716$ and AR$= 0.5454$. Likewise, the three CIFAR-10 settings have prediction agreement 0.9561–0.9616 but AR only 0.5849–0.6238. Civil is substantially more reproducible, with entropy Spearman 0.9723, prediction agreement 0.9908, and AR$= 0.8868$.

Raw AR generally rises with review capacity, but so does its chance baseline (Table 2). For example, from 20% to 30%, R18/C10 raw AR rises from 0.7024 to 0.7118, while chance-corrected AR falls from 0.6280 to 0.5883. Full adjusted results are in Appendix D.

Selective utility also varies across runs. At 10% review, retained-risk ranges are 0.267–1.711 percentage points for vision and 0.055 for Civil; actual error-coverage ranges are 1.345–6.549 and 1.192 percentage points, respectively. Full results and area under the risk–coverage curve (AURC) are reported in Appendix D. No practical stability tolerance was prespecified.

Table 1: Top-10% results across ten retrainings. Entropy ρ is mean pairwise average-tie Spearman correlation, and AR is normalized review-set overlap. Common-error metrics use $q = 8$. Accuracy is mean $\pm$ standard deviation in percent; Δ_{ECR} is the across-run range of common-error coverage.

Setting	Accuracy (%)	Pred. agr.	Entropy ρ	AR	CEFR	PairDis	Δ_{ECR}
R18/C10	95.03 ± 0.14	0.9573	0.6818	0.5981	0.8199	0.3309	0.0651
R18/C100	77.38 ± 0.22	0.8134	0.8452	0.5486	0.5875	0.2326	0.0274
R34/C10	95.16 ± 0.15	0.9561	0.6736	0.5849	0.8397	0.3245	0.0633
R34/C100	77.97 ± 0.36	0.8151	0.8162	0.4994	0.6523	0.2683	0.0284
D121/C10	95.61 ± 0.11	0.9616	0.6377	0.6238	0.8238	0.3211	0.0810
R18/ImageNet-100	73.85 ± 0.50	0.7770	0.8716	0.5454	0.6065	0.2439	0.0346
DBERT/Civil	95.53 ± 0.03	0.9908	0.9723	0.8868	0.2648	0.0959	0.0126

Table 2: Mean pairwise AR across review budgets and ten retrainings.

Setting	1%	5%	10%	20%	30%
R18/C10	0.2247	0.4166	0.5981	0.7024	0.7118
R18/C100	0.2200	0.4281	0.5486	0.6736	0.7384
R34/C10	0.2153	0.3972	0.5849	0.6883	0.6954
R34/C100	0.1624	0.3733	0.4994	0.6367	0.7158
D121/C10	0.2407	0.4323	0.6238	0.6837	0.6785
R18/ImageNet-100	0.2529	0.4228	0.5454	0.6821	0.7533
DBERT/Civil	0.3521	0.8082	0.8868	0.9155	0.9218

5.3 Instability Reaches Persistent Errors

At top-10%, 58.75–83.97% of visual common errors ($q = 8$) change review status across retrainings, with PairDis 0.2326–0.3309. At 1% review, 55–82% of CIFAR common errors are never selected, and none are always selected, showing that a low any-run flip rate can coexist with limited common-error coverage. At 10%, only 2.36–12.38% are always reviewed. For ImageNet-100, 35.01% of 897 common errors are never reviewed, 60.65% flip, and 4.35% are always reviewed.

Civil is more stable: among 3,822 common errors, CEFR$= 0.2648$ and PairDis$= 0.0959$; 50.73% are always reviewed and 22.79% never reviewed. Among 2,487 common toxic false negatives, 27.26% change review status. Full budget-wise partitions are reported in Appendix C.3, with additional Civil metrics in Appendix C.5.

Allocation and prediction churn are not equivalent. Among high-instability cases—selected by four to six of ten runs at top-10%, equivalently $\widehat{\mathrm{IDI}}_i \geq 0.96$—mean pairwise class disagreement is 0.0007 for Civil and 0.1631–0.6393 across the visual settings. Review status also changes among cases whose predicted class is constant across all ten runs: 1,156/8,903 in R18/C10, 101/2,846 in ImageNet-100, and 5,933/94,875 in Civil. Thus allocation turnover cannot be inferred from prediction churn alone. Additional diagnostics are in Appendix C.12.

5.4 Boundary Diagnostics and Certificate Evaluation

Across all $7 \times 5 \times 45 = 1{,}575$ setting–budget–pair cases, Proposition 3 has zero violations. The bound is nontrivial in 1,519 cases (96.44%), with mean clipped slack 0.0583; all 56 trivial cases occur at the 1% budget (Appendix A.8). Here nontrivial means that the turnover bound is below one, and clipped slack is the clipped bound minus observed turnover.

Cutoff proximity has a setting-dependent association with empirical IDI (Table 3). The Gaussian surrogate, which additionally uses across-run score variance, has numerically higher correlations in all seven settings (0.5007–0.7829). Because these correlations are computed from the same retraining ensemble, we treat them as descriptive. Additional cutoff-distance diagnostics and the standardized Gaussian interpretation are provided in Appendices C.4 and A.6.

The uniform ℓ_∞ guarantees are substantially more conservative: the exact-radius condition certifies none of the 1,575 observed cases, and the multi-swap bound is vacuous for all. Adjacent order-statistic

Table 3: Boundary diagnostics at top-10%. The first column correlates empirical IDI with negative distance between the across-run mean score μ_i and mean cutoff τ. The second correlates empirical IDI with the Gaussian surrogate $\widehat{\mathrm{IDI}}^{G} = 4\Phi(z)(1-\Phi(z))$, where $z = (\mu_i - \tau)/\sigma_i$.

Setting	$\rho(-\lvert\mu-\tau\rvert, \widehat{\mathrm{IDI}})$	$\rho(\widehat{\mathrm{IDI}}, \widehat{\mathrm{IDI}}^{G})$
R18/C10	0.2768	0.7677
R18/C100	0.7218	0.7543
R34/C10	0.0979	0.7825
R34/C100	0.7460	0.7829
D121/C10	0.1264	0.7249
R18/ImageNet-100	0.7173	0.7551
DBERT/Civil	0.4124	0.5007

gaps near the cutoff can be small, whereas $\|u - v\|_\infty$ reflects the largest score displacement over the full evaluation set.

5.5 Robustness Checks and Secondary Diagnostics

The main pattern persists across $q \in \{6, 8, 10\}$ and when predictive entropy is replaced by $1 - \mathrm{MSP}$, with MSP denoting maximum softmax probability, although allocation can depend on the uncertainty score. For ImageNet-100, top-10% AR changes from 0.5454 to 0.4424, with policy overlap 0.8040; entropy and $1 - \mathrm{MSP}$ induce the same ordering for binary Civil classification (Appendices C.9 and C.10). Alternative deterministic tie-breaking rules change Civil AR by less than 1.2×10^{-4} across budgets (Appendix C.11).

Initialization and training controls. On R18/C10, top-10% AR is 0.5981 with independent initialization, 0.5961 from a shared random initialization, and 0.5383 from a shared ImageNet-pretrained initialization; corresponding accuracies are 95.03%, 95.00%, and 94.27%. Shared initialization therefore does not reproduce Civil's high AR in this R18/C10 control. Because the pretrained condition also has lower accuracy, this comparison does not isolate the effect of pretraining (Appendix C.7).

Longer training improves visual accuracy by 1.95–5.53 percentage points but does not consistently improve AR; for example, R34/C100 decreases from 0.5238 to 0.4994 at top-10%. A separate three-seed Civil comparison supports the two-epoch protocol used in the main experiment (Appendix C.6).

A capacity-calibrated Gaussian surrogate tracks AR across the 35 setting–budget cells with Spearman 0.9821 and MAE 0.0427. Because budgets are dependent within settings and both surrogate statistics and empirical AR are computed from the same retrainings, this is a descriptive same-ensemble association rather than an out-of-sample prediction result (Appendix C.8).

Probability averaging. We evaluate whether standard probability averaging reduces allocation turnover without degrading predictive performance. Each of the 945 perfect matchings partitions the ten runs into five two-model ensembles. Within each pair, predictive probabilities are averaged before recomputing entropy and applying the unchanged exact top-k policy.

Table 4 reports the top-10% results. The median matching-level ΔAR is positive in all seven settings and in all 35 setting–budget cells. Mean two-model accuracy, computed over all 45 unordered run pairs, is also higher than mean single-model accuracy in every setting. Standard probability averaging reduces, but does not eliminate, allocation turnover, at the cost of two-model inference. The 945 matchings are not independent replicates.

Retraining-count sensitivity. Enumerating all subsets of $S' \in \{2, 3, 4, 5, 6, 8\}$ runs gives a median absolute deviation of 0.0008–0.0055 from the ten-run AR reference at $S' = 5$. This analysis only measures sensitivity within the observed ten-run ensemble; it does not imply that five runs are sufficient in general (Appendix C.2).

Table 4: Two-model probability averaging at 10% review. $\Delta\mathrm{AR}$ is the median matching-level improvement over the single-model baseline. Two-model accuracy is the mean over all 45 unordered run pairs.

Setting	Baseline AR	Median ΔAR	Accuracy (%)
R18/C10	0.5981	+0.0847	95.03→95.59
R18/C100	0.5486	+0.1146	77.38→79.29
R34/C10	0.5849	+0.0820	95.16→95.80
R34/C100	0.4994	+0.1266	77.97→79.80
D121/C10	0.6238	+0.0805	95.61→96.12
R18/ImageNet-100	0.5454	+0.1132	73.85→75.94
DBERT/Civil	0.8868	+0.0332	95.53→95.55

6 Limitations and Scope

Our experiments cover seven model–dataset settings with ten retrainings each. ImageNet-100 extends the evaluation beyond CIFAR, but remains a fixed 100-class subset. We study post-hoc exact top-k review; learned rejectors and variable-capacity policies may behave differently.

The vision–Civil comparison also differs in architecture, pretraining, and optimization. Our R18/C10 control examines initialization, but does not separate these factors completely. The boundary analysis concerns the score changes observed after retraining rather than their optimization-level cause.

Allocation turnover measures changes in who is reviewed, not the downstream cost of those changes. Probability averaging reduces turnover in our experiments, but requires two-model inference and has not been evaluated outside the settings considered here.

7 Conclusion

Independent retraining can change review assignments despite small variation in accuracy. At a 10% review budget, mean pairwise turnover is 37.6–50.1% across the six visual settings and 11.3% on Civil Comments. Its extent varies across settings, and aggregate performance or global agreement alone does not certify preservation of review assignments.

We derive a sharp rank–allocation bound, use an established subset-stability identity to connect set-level and case-level disagreement, and bound membership changes through score and cutoff perturbations. Two-model probability averaging reduces median matching-level turnover in every tested setting and budget, but leaves residual disagreement. When review identity matters, allocation reproducibility captures information that predictive performance and selective utility do not.

Reproducibility Statement

All seven main settings use ten independently randomized retrainings (seeds 1–10). Pairwise quantities are computed over all 45 unordered pairs of runs. Because these pairs share retrainings, we do not treat them as independent replicates; uncertainty is estimated with a delete-one-retraining jackknife.

Review sets are constructed with exact $k = \lceil \alpha n \rceil$ selection, ranking examples by descending uncertainty and breaking ties by ascending sample index. The primary common-error analysis uses $q = 8$, with $q \in \{6, 8, 10\}$ used for sensitivity analysis. Robustness to the uncertainty score is evaluated in all seven settings by replacing predictive entropy with $1 - \mathrm{MSP}$ while keeping the same final probability outputs and selection procedure.

The retraining-count and two-model analyses are performed from stored final probability outputs. For retraining-count sensitivity, we enumerate all $\binom{10}{S}$ subsets for $S \in \{2, 3, 4, 5, 6, 8\}$. For the two-model analysis, we enumerate all 945 perfect matchings of the ten retrainings, average probabilities within each pair, recompute entropy, and apply the same top-k rule and tie-breaking procedure. Training details, implementation settings, and additional robustness results are given in the appendix.

AI USE STATEMENT

Generative AI tools were used to assist with language revision and polishing, retrieval and discovery of related work, and checking mathematical derivations and proofs. The authors reviewed and verified all AI-assisted content and are responsible for the claims, proofs, references, and results in the paper.

REFERENCES

Dara Bahri and Heinrich Jiang. Locally adaptive label smoothing improves predictive churn. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 532–542. PMLR, 2021.

Emily Black, Klas Leino, and Matt Fredrikson. Selective ensembles for consistent predictions. In *International Conference on Learning Representations*, 2022.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500. ACM, 2019. doi: 10.1145/3308560.3317593.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769, 2021.

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

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, pages 2902–2913, 2019.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003. doi: 10.1137/S0895480102412856.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

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*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 3207–3214, 2018. doi: 10.1609/aaai.v32i1.11694.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Max Hopkins and Shay Moran. The role of randomness in stability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 23805–23827. PMLR, 2025.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. doi: 10.1109/CVPR.2017.243.

Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 818–831, 2022. doi: 10.1145/3519935.3519973.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pages 6402–6413, 2017.

Ruiting Liang, Jake A. Soloff, Rina Foygel Barber, and Rebecca Willett. Assumption-free stability for ranking problems, 2025.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6765–6774. PMLR, 2020.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. doi: 10.1111/j.1467-9868.2010.00740.x.

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174):1–54, 2018.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 13991–14002, 2019.

Marios Papamichalis, Regina Ruane, and Theofanis Papamichalis. Replicable conformal prediction, 2026. Preprint.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter, 2019. Accepted at the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, NeurIPS 2019.

Cecilia Summers and Michael J. Dinneen. Nondeterminism and instability in neural network optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9913–9922. PMLR, 2021.

A PROOFS

A.1 PROOF OF THEOREM 1

Relabel items by their ranks under π, so that π is the identity. This preserves both the squared-rank sum and top-k overlap. Let $X = \{1, \dots, k\} \setminus S_\sigma^k$ and $Y = S_\sigma^k \setminus \{1, \dots, k\}$, each of size r. The case $r = 0$ is immediate; assume $r \geq 1$.

Write $X = \{x_1 < \cdots < x_r\}$ and let $t_1 < \cdots < t_r$ be its image ranks under σ. Then $x_a \leq k - r + a$ and $t_a \geq k + a$. Order-preserving matching minimizes squared displacement: swapping an inverted assignment $x_a < x_b$ to $t_c > t_d$ reduces its cost by $2(x_b - x_a)(t_c - t_d) > 0$. Therefore

$$\sum_{i \in X} (i - \sigma_i)^2 \geq \sum_{a=1}^{r} (t_a - x_a)^2 \geq r^3.$$

The same argument for Y, whose source ranks lie above k and target ranks at most k, gives another r^3. Since X and Y are disjoint, $\sum_i (i - \sigma_i)^2 \geq 2r^3$.

To attain equality, exchange the adjacent blocks $\{k-r+1, \dots, k\}$ and $\{k+1, \dots, k+r\}$, preserving order within each block and fixing all other items. Exactly $2r$ items move r positions, producing r replacements and squared displacement $2r^3$. The Spearman bound follows from

$$\rho = 1 - \frac{6 \sum_i (\pi_i - \sigma_i)^2}{N(N^2 - 1)}.$$

Identical rankings attain equality for $r = 0$. □

Corollary 1 follows by rearranging this bound and using $r \in \{0, \dots, \min(k, N - k)\}$. □

A.2 PROOF OF PROPOSITION 1

Let $a = |A \cap B|$. Set inclusion gives $a \leq \min(r_1, r_2)$, while $|A \cup B| = r_1 + r_2 - a \leq m$ gives $a \geq \max(0, r_1 + r_2 - m)$. For any integer a in this interval, choose a shared elements, $r_1 - a$ elements unique to A, and $r_2 - a$ unique to B. This uses at most m elements, proving attainability. For $r_1 = r_2 = r$ with $1 \leq r \leq m/2$, the attainable normalized overlaps are $\{0, 1/r, \dots, 1\}$ despite equal coverage. □

A.3 FINITE-RETRAINING BIAS AND ALLOCATION DECOMPOSITION

Proposition 4 (Finite-retraining bias of empirical IDI)**.** *For $S \geq 2$, suppose $Z_{i,1}, \dots, Z_{i,S}$ are i.i.d.* $\mathrm{Bernoulli}(p_i)$ *and let $\hat{p}_i = S^{-1} \sum_s Z_{i,s}$. Then*

$$\mathbb{E}[4\hat{p}_i(1 - \hat{p}_i)] = \frac{S - 1}{S} 4p_i(1 - p_i).$$

Thus the corrected estimator

$$\widetilde{\mathrm{IDI}}_i = \frac{4S}{S - 1} \hat{p}_i(1 - \hat{p}_i)$$

is unbiased for $4p_i(1 - p_i)$.

Proof. Since $\mathbb{E}[\hat{p}_i] = p_i$ and $\mathrm{Var}(\hat{p}_i) = p_i(1 - p_i)/S$,

$$\mathbb{E}[\hat{p}_i(1 - \hat{p}_i)] = p_i(1 - p_i) - \mathrm{Var}(\hat{p}_i) = \frac{S - 1}{S} p_i(1 - p_i).$$

Multiplication by four proves the claim. □

Proof of Proposition 2. Case i disagrees on $m_i(S - m_i)$ unordered run pairs. As every set has size k,

$$\sum_i D_i = \binom{S}{2}^{-1} \sum_{s<t} |\mathcal{S}_s \triangle \mathcal{S}_t| = 2\binom{S}{2}^{-1} \sum_{s<t} \big(k - |\mathcal{S}_s \cap \mathcal{S}_t|\big) = 2k(1 - \mathrm{AR}).$$

□

Corollary 2 (Relation to corrected IDI). *If each of $S \geq 2$ runs selects exactly $k \geq 1$ cases from the same universe, then*

$$1 - \mathrm{AR} = \frac{S}{(S-1)k} \sum_i \hat{p}_i(1-\hat{p}_i), \qquad \sum_i \widetilde{\mathrm{IDI}}_i = 4k(1-\mathrm{AR}).$$

Proof. Substitute $D_i = 2S\hat{p}_i(1-\hat{p}_i)/(S-1)$ and $\widetilde{\mathrm{IDI}}_i = 2D_i$ into the preceding decomposition. □

These identities hold deterministically for the observed runs. Independence is needed only for the unbiasedness interpretation in Proposition 4.

A.4 Aggregation and covariance cancellation

For a fixed subpopulation of size m, let $A_s = m^{-1}\sum_{i=1}^{m} Z_{i,s}$. Variance over retraining satisfies

$$\mathrm{Var}(A_s) = \frac{1}{m^2}\left[\sum_i \mathrm{Var}(Z_i) + 2\sum_{i<j} \mathrm{Cov}(Z_i, Z_j)\right]. \tag{22}$$

Proposition 5 (Aggregate masking). *Consider a sequence of such populations as $m \to \infty$. If $m^{-1}\sum_i \mathrm{Var}(Z_i) \geq c$ for a constant $c > 0$ independent of m, and $\sum_{i<j} \mathrm{Cov}(Z_i, Z_j) = O(m)$, then average individual variance is $\Omega(1)$ whereas $\mathrm{Var}(A_s) = O(m^{-1})$. For even m, exact cancellation is possible: let $Z_{2j,s} = 1 - Z_{2j-1,s}$ with $Z_{2j-1,s} \sim \mathrm{Bernoulli}(1/2)$. Then every case has variance $1/4$, but $A_s \equiv 1/2$.*

Proof. Bernoulli variances are at most $1/4$, so $\sum_i \mathrm{Var}(Z_i) = O(m)$. The covariance assumption and Eq. equation 22 give the aggregate bound; the individual bound is assumed. In the construction, every pair sums to one. □ The construction shows that such cancellation is possible; we do not claim that it explains the empirical results.

A.5 Sharp deterministic stability results

Let $u, v \in \mathbb{R}^n$. Write $u_{(1)} \geq \cdots \geq u_{(n)}$, using the fixed sample-index rule to order ties, and let S_u, S_v denote the corresponding exact top-k sets.

Theorem 2 (Allocation-change profile). *For $1 \leq k < n$ and $1 \leq t \leq \min(k, n-k)$, define $G_t(u) = u_{(k-t+1)} - u_{(k+t)}$. Then*

$$\|u - v\|_\infty < G_t(u)/2 \quad \Longrightarrow \quad k - |S_u \cap S_v| \leq t - 1.$$

For every $\varepsilon > 0$, a perturbation of norm at most $G_t(u)/2 + \varepsilon$ can cause at least t replacements.

Corollary 3 (Sharp certified top-k radius). *For $1 \leq k < n$, let $\gamma_k(u) = u_{(k)} - u_{(k+1)}$. If $\|u - v\|_\infty < \gamma_k(u)/2$, then $S_u = S_v$. For every $\varepsilon > 0$, some v satisfies*

$$\|u - v\|_\infty \leq \gamma_k(u)/2 + \varepsilon, \qquad S_u \neq S_v.$$

Thus $\gamma_k(u)/2$ is the supremal uniform radius for the open perturbation ball. For $k = n$, the selected set is invariant.

Proof of Theorem 2. If at least t items are replaced, some removed item i has original rank at most $k - t + 1$, and some entering item j has original rank at least $k + t$. Thus $u_i - u_j \geq G_t(u)$ and $v_j \geq v_i$, giving

$$G_t(u) \leq u_i - u_j \leq (u_i - v_i) + (v_j - u_j) \leq 2\|u - v\|_\infty.$$

This contradicts the strict norm bound. For sharpness, lower the bottom t selected scores and raise the top t unselected scores by $G_t(u)/2 + \varepsilon$. Every raised score strictly exceeds every lowered score and all unchanged unselected scores. At most $k - t$ other items can precede a raised item, so all t raised items enter the top-k set. □

Proof of corollary 3. Take $t = 1$ in Theorem 2, since $G_1(u) = \gamma_k(u)$. For $k = n$, both sets are the full universe. □

A.6 Boundary-proximity diagnostic under the Gaussian approximation

For $\sigma_i > 0$, define $z_i = (\mu_i - \tau)/\sigma_i$, $p_i^G = \Phi(z_i)$, and $\widehat{\mathrm{IDI}}_i^G = 4p_i^G(1 - p_i^G)$, where Φ and ϕ are the standard normal CDF and density. For $\sigma_i = 0$, set $p_i^G = 1$ if $\mu_i > \tau$, 0 if $\mu_i < \tau$, and $1/2$ at equality. The equality case is a surrogate convention, not the allocation tie-breaking rule. The function $f(z) = 4\Phi(z)(1 - \Phi(z))$ is even and satisfies

$$f'(z) = 4\phi(z)(1 - 2\Phi(z)).$$

It therefore decreases strictly with $|z|$ and attains its maximum $f(0) = 1$. This concerns standardized cutoff distance; monotonicity in $|\mu_i - \tau|$ alone requires fixed $\sigma_i > 0$.

A.7 Proof of Proposition 3

For $i \in S_u \triangle S_v$, let $a_i = u_i - \tau_u$ and $b_i = v_i - \tau_v$. Membership change implies $a_i b_i \leq 0$, including ties. Hence

$$\max\{|a_i|, |b_i|\} \leq |a_i - b_i| \leq |u_i - v_i| + |\tau_u - \tau_v|,$$

which proves $S_u \triangle S_v \subseteq R_u \cap R_v$. The overlap bounds follow from $|S_u \triangle S_v| = 2(k - |S_u \cap S_v|)$ and $|R_u \cap R_v| \leq \min(|R_u|, |R_v|)$. If $\|u - v\|_\infty \leq \delta$, the kth order statistic is 1-Lipschitz under ℓ_∞, so $|\tau_u - \tau_v| \leq \delta$. The preceding inequality then places each changed item within 2δ of both corresponding cutoffs. For $k = n$, the change set is empty. □

A.8 Certificate Evaluation Details

For each run pair, let $B = |R_u \cap R_v|/(2k)$ and $T = 1 - |S_u \cap S_v|/k$. The bound is nontrivial when $B < 1$; its clipped slack is $\min(1, B) - T$. Across the 1,575 setting–budget–pair cases, 1,519 (96.44%) have nontrivial bounds, and mean clipped slack is 0.0583. All 56 trivial bounds occur at the 1% budget: 259/315 bounds are nontrivial there, versus 1,260/1,260 at larger budgets. For R18/ImageNet-100, the corresponding counts are 44/45 at 1% and 45/45 at each larger budget.

B Experimental Details

Image classification. We use the official CIFAR-10 and CIFAR-100 training/test splits with CIFAR-adapted ResNet-18, ResNet-34, and DenseNet-121 models. The initial convolution is replaced by a 3×3 stride-1 layer and the initial max-pooling layer is removed. Models are trained from scratch with cross-entropy, SGD (learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}), batch size 128, cosine decay without warmup, random 32×32 crops with 4-pixel padding, and horizontal flipping. R18/C10, R18/C100, R34/C10, and D121/C10 use 100 epochs; R34/C100 uses a predefined 110-epoch schedule. Each setting uses seeds 1–10.

ImageNet-100 replication. We use a fixed 100-class ImageNet subset (Deng et al., 2009). From a 130,000-image class-balanced pool, we select 500 images per class for training (50,000 total) and evaluate on a disjoint 5,000-image query set (50 per class). R18 uses the standard ImageNet stem and 224×224 inputs. Training starts from random initialization and runs for 50 epochs with cross-entropy, SGD (learning rate 0.1, momentum 0.9, weight decay 10^{-4}), batch size 256, five-epoch linear warmup followed by cosine decay, and mixed precision. Training uses random resized crops and horizontal flipping; evaluation uses resize to 256 followed by a 224×224 center crop. Seeds 1–10 use the same predefined epoch-50 endpoint.

Civil Comments. We fine-tune `distilbert-base-uncased` on fixed train, validation, and test splits containing 1,804,874, 97,320, and 97,320 examples, respectively. Toxicity scores of at least 0.5 define the positive class, and the full test split is used for allocation analysis. Ten retrainings (seeds 1–10) use two fine-tuning epochs, AdamW with learning rate 2×10^{-5} and weight decay 0.01, batch size 64, maximum sequence length 128, 10% linear warmup followed by linear decay, gradient clipping at 1.0, and mixed precision. We use no class weighting, oversampling, early stopping, or seed-specific tuning.

Table 5: Delete-one-retraining jackknife standard errors at the top-10% review budget. Point estimates use all ten retrainings.

Setting	Pred. agr.	JK SE	Entropy ρ	JK SE	AR@10%	JK SE
R18/C10	0.9573	0.000586	0.6818	0.005111	0.5981	0.004375
R18/C100	0.8134	0.001011	0.8452	0.000734	0.5486	0.004061
R34/C10	0.9561	0.000668	0.6736	0.018475	0.5849	0.004564
R34/C100	0.8151	0.002189	0.8162	0.003770	0.4994	0.003354
D121/C10	0.9616	0.000498	0.6377	0.004506	0.6238	0.003985
R18/ImageNet-100	0.7770	0.0016	0.8716	0.0012	0.5454	0.0051
DBERT/Civil	0.9908	0.000070	0.9723	0.000765	0.8868	0.001173

Post-hoc analysis. Review sets use predictive entropy and exact capacity $k = \lceil \alpha n \rceil$ at $\alpha \in \{1, 5, 10, 20, 30\}\%$, with the fixed sample-index tie-breaking rule. Common-error analyses use $q = 8$ as primary and $q \in \{6, 8, 10\}$ for sensitivity. Across-run standard deviations and variances use the population convention (divisor S). Pairwise metrics use all 45 unordered retraining pairs; uncertainty is estimated by delete-one-retraining jackknife. For Civil Comments, alternative deterministic tie-breaking rules change AR by at most 1.2×10^{-4} across the evaluated budgets.

C Supplementary Analyses

C.1 Jackknife uncertainty

Table 5 reports delete-one-retraining jackknife standard errors for the fixed-universe pairwise metrics at the top-10% review budget. Each jackknife replicate omits one retraining and recomputes the statistic from the remaining 36 unordered pairs. For a statistic $\hat{T}$, let $\hat{T}_{(-s)}$ denote its estimate after omitting run s, and let $\bar{T}_{(-\cdot)} = S^{-1} \sum_s \hat{T}_{(-s)}$. We report

$$\widehat{\mathrm{SE}}_{\mathrm{JK}} = \left[\frac{S-1}{S} \sum_s \big(\hat{T}_{(-s)} - \bar{T}_{(-\cdot)} \big)^2 \right]^{1/2}.$$

These standard errors quantify retraining variation conditional on the fixed evaluation population and therefore do not include test-set sampling uncertainty. CEFR and PairDis are not assigned these order-2 U-statistic standard errors because $\mathcal{E}_q$ is constructed jointly from all S retrainings.

C.2 Two-model probability averaging and retraining-count sensitivity

Each perfect matching partitions the ten retrainings into five disjoint unordered pairs. Within each pair, predictive probability vectors are averaged before recomputing entropy and applying the same exact top-k rule. The number of such matchings is

$$\frac{10!}{2^5 5!} = 9!! = 945.$$

Because all matchings reuse the same ten runs, they are not independent replicates. For a matching M, let $\mathrm{AR}_M^{(2)}$ denote the mean normalized overlap over the $\binom{5}{2} = 10$ unordered pairs of its five two-model ensembles, and define

$$\Delta\mathrm{AR}_M = \mathrm{AR}_M^{(2)} - \mathrm{AR}_{\mathrm{base}},$$

where $\mathrm{AR}_{\mathrm{base}}$ is the original ten-run single-model estimate. We report the median of $\Delta\mathrm{AR}_M$ over the 945 matchings.

For each unordered pair of retrainings, accuracy is computed from the argmax of the averaged probabilities. Reported two-model accuracy is the arithmetic mean over all 45 unordered pairs. Equivalently, it is the mean of the five-pair averages over all 945 matchings, since every unordered pair occurs in exactly 105 matchings.

Table 6 summarizes the top-10% results. Median matching-level $\Delta\mathrm{AR}$ is positive in all seven settings at 10% review and in all 35 setting–budget cells. Mean two-model accuracy also exceeds mean

Table 6: Two-model probability averaging at the top-10% review budget. Baseline AR is the ten-run single-model estimate, changed slots equals $k(1 - \mathrm{AR})$, and $\Delta\mathrm{AR}$ is the median matching-level improvement. Accuracy reports the single-model mean followed by the mean over all 45 two-model pairs.

Setting	Baseline AR	Changed slots	Median ΔAR	Accuracy (%)
R18/C10	0.5981	401.9 / 1000	+0.0847	95.03→95.59
R18/C100	0.5486	451.4 / 1000	+0.1146	77.38→79.29
R34/C10	0.5849	415.1 / 1000	+0.0820	95.16→95.80
R34/C100	0.4994	500.6 / 1000	+0.1266	77.97→79.80
D121/C10	0.6238	376.2 / 1000	+0.0805	95.61→96.12
R18/ImageNet-100	0.5454	227.3 / 500	+0.1132	73.85→75.94
DBERT/Civil	0.8868	1101.3 / 9732	+0.0332	95.53→95.55

single-model accuracy in every setting. Thus averaging improves median allocation reproducibility in all evaluated cells and mean accuracy in all settings; deployment requires two-model inference.

Retraining-count sensitivity. For $S' \in \{2, 3, 4, 5, 6, 8\}$, we enumerate all $\binom{10}{S'}$ subsets and recompute AR, using the ten-run estimate as the reference. At $S' = 5$, the median absolute deviation ranges from 0.0008 for Civil to 0.0055 for R34/C10. This quantifies finite-run sensitivity for the evaluated settings and does not imply that five retrainings are generally sufficient.

C.3 Common-error instability

For each common error, let $m_i = \sum_{s=1}^{S} Z_{i,s}$ be the number of retrainings in which it is selected for review. We partition the common-error universe into never-reviewed ($m_i = 0$), flipping ($0 < m_i < S$), and always-reviewed ($m_i = S$) cases; the flipping fraction equals CEFR. Table 7 reports this decomposition. At 1% review, no common error is always selected in the five CIFAR settings or Civil, while the never-reviewed fraction is 0.5527–0.8211 for CIFAR and 0.7493 for Civil. Thus a low flip rate can coexist with limited common-error coverage. At 10%, visual common errors still show substantial reassignment: 0.5875–0.8397 flip review status, while only 0.0236–0.1238 are always reviewed. Civil is more stable, with 0.5073 always reviewed and 0.2648 flipping.

C.4 Sample-level stability diagnostics

Table 8 reports cutoff-distance diagnostics for the six settings with complete distance summaries. The high-instability subset consists of examples selected by four to six of the ten retrainings, equivalently $\widehat{\mathrm{IDI}}_i \geq 0.96$.

High-instability examples lie closer to the cutoff than always-selected examples in all six settings. Their distance relative to never-selected examples varies across settings, so cutoff distance alone does not yield a consistent ordering of case-level instability.

C.5 Civil Comments detailed results

Table 9 separates the full $\mathcal{E}_8$ common-error universe from the toxic false-negative subset used in the targeted analysis.

C.6 Training-maturity robustness

Table 10 compares shorter-duration and final configurations for the six visual settings. The final configurations have higher accuracy in all six visual settings, while changes in AR are heterogeneous. A separate three-seed Civil comparison gives validation AUROC of 0.9660 ± 0.0001, 0.9628 ± 0.0001, and 0.9576 ± 0.0003 after two, three, and four epochs, respectively. Among these durations, two epochs gives the highest mean validation AUROC.

Table 7: Decomposition of the $q = 8$ common-error universe across ten retrainings. R18/ImageNet-100 is reported at the primary 10% budget. Never, Flip, and Always denote review in none, some, or all retrainings; Flip equals CEFR. Mean coverage is common-error coverage averaged across runs; coverage range is its maximum minus minimum across runs.

Setting	Budget	$\lvert\mathcal{E}_8\rvert$	Never	Flip (CEFR)	Always	Mean coverage	Coverage range	PairDis
R18/C10	1%	261	0.5632	0.4368	0.0000	0.1100	0.0192	0.1557
R18/C10	5%	261	0.1188	0.8659	0.0153	0.3828	0.0575	0.3665
R18/C10	10%	261	0.0613	0.8199	0.1188	0.6192	0.0651	0.3309
R18/C100	1%	1532	0.8211	0.1789	0.0000	0.0392	0.0072	0.0580
R18/C100	5%	1532	0.5320	0.4582	0.0098	0.1765	0.0189	0.1761
R18/C100	10%	1532	0.3642	0.5875	0.0483	0.3185	0.0274	0.2326
R34/C10	1%	237	0.5527	0.4473	0.0000	0.1063	0.0295	0.1571
R34/C10	5%	237	0.1266	0.8608	0.0127	0.3785	0.0844	0.3675
R34/C10	10%	237	0.0591	0.8397	0.1013	0.6241	0.0633	0.3245
R34/C100	1%	1481	0.7745	0.2255	0.0000	0.0422	0.0068	0.0682
R34/C100	5%	1481	0.4733	0.5233	0.0034	0.1847	0.0149	0.2064
R34/C100	10%	1481	0.3241	0.6523	0.0236	0.3313	0.0284	0.2683
D121/C10	1%	210	0.5714	0.4286	0.0000	0.1076	0.0571	0.1534
D121/C10	5%	210	0.1571	0.8381	0.0048	0.3629	0.0619	0.3558
D121/C10	10%	210	0.0524	0.8238	0.1238	0.6057	0.0810	0.3211
R18/ImageNet-100	10%	897	0.3501	0.6065	0.0435	0.3253	0.0346	0.2439
DBERT/Civil	1%	3822	0.7493	0.2507	0.0000	0.0622	0.0123	0.0885
DBERT/Civil	5%	3822	0.4409	0.3851	0.1740	0.3748	0.0118	0.1430
DBERT/Civil	10%	3822	0.2279	0.2648	0.5073	0.6570	0.0126	0.0959

Table 8: Top-10% sample-level stability diagnostics. Distances are median $|\mu_i - \tau|$ within the high-instability, always-selected, and never-selected groups, where μ_i is mean entropy across retrainings and τ is the mean top-k cutoff.

Setting	High-inst. frac.	Cutoff mean	Cutoff std	Unstable dist.	Always dist.	Never dist.
R18/C10	0.0547	0.1632	0.0158	0.1099	0.6000	0.1588
R18/C100	0.0640	2.2700	0.0257	0.1869	0.9454	2.1827
R34/C10	0.0577	0.0927	0.0165	0.1192	0.5264	0.0903
R34/C100	0.0733	1.7821	0.0466	0.1827	1.0356	1.7399
D121/C10	0.0470	0.1030	0.0059	0.1219	0.5394	0.0987
DBERT/Civil	0.0150	0.3678	0.0058	0.0185	0.2623	0.3655

Table 9: Civil Comments/DistilBERT results across ten retrainings. Performance metrics are mean $\pm$ standard deviation.

Metric	Result
AUROC	0.9666 ± 0.0002
Average Precision	0.7930 ± 0.0008
Accuracy	0.9553 ± 0.0003
Prediction agreement	0.9908
Entropy Spearman	0.9723
AR 1/5/10%	0.3521 / 0.8082 / 0.8868
CEFR 1/5/10%	0.2507 / 0.3851 / 0.2648
PairDis 1/5/10%	0.0885 / 0.1430 / 0.0959
ECR range 1/5/10%	0.0123 / 0.0118 / 0.0126
Full common-error universe $\lvert\mathcal{E}_8\rvert$	3822
Common toxic false negatives ($q = 8$)	2487
Toxic-FN flip rate	0.2726

C.7 Initialization-regime control

We compare the R18/C10 random-scratch baseline with two same-task controls: ten retrainings from a shared random initialization and ten from a shared ImageNet-pretrained backbone. The downstream

Table 10: Training-maturity comparison for the six visual settings.

Setting	Short ep.	Final ep.	Short acc.	Final acc.	Δ acc.	Short AR@10%	Final AR@10%
R18/C10	20	100	92.08	95.03	+2.95	0.5328	0.5981
R18/C100	20	100	73.12	77.38	+4.26	0.5441	0.5486
R34/C10	20	100	91.60	95.16	+3.56	0.4982	0.5849
R34/C100	20	110	72.78	77.97	+5.19	0.5238	0.4994
D121/C10	20	100	93.66	95.61	+1.95	0.5706	0.6238
R18/ImageNet-100	30	50	68.32	73.85	+5.530	0.4624	0.5454

CIFAR training protocol, evaluation set, and allocation procedure are held fixed. Table 11 reports the results.

Table 11: R18/C10 initialization-regime control. Accuracy is mean $\pm$ standard deviation across ten retrainings; all rows use the same downstream training and exact top-k evaluation protocol.

Initialization	Accuracy (%)	Pred. agr.	AR@1%	AR@5%	AR@10%	AR@20%	AR@30%
Random scratch	95.03 ± 0.135	0.9573	0.2247	0.4166	0.5981	0.7024	0.7118
Shared random	95.00 ± 0.134	0.9567	0.2409	0.4116	0.5961	0.6985	0.7119
Shared ImageNet	94.27 ± 0.320	0.9483	0.1978	0.3745	0.5383	0.6846	0.7159

Shared random initialization leaves top-10% AR nearly unchanged (0.5981 versus 0.5961) and does not reproduce Civil's substantially higher reproducibility. The shared-ImageNet condition also yields lower AR (0.5383), but its lower accuracy prevents attributing this difference to pretraining alone.

C.8 Gaussian score surrogate

We approximate marginal review probabilities using each example's across-retraining entropy mean μ_i and standard deviation σ_i. For a cutoff t, define $p_i^G(t) = \Phi((\mu_i - t)/\sigma_i)$ when $\sigma_i > 0$. When $\sigma_i = 0$, set $p_i^G(t)$ to 1, 0, or $1/2$ according as $\mu_i > t$, $\mu_i < t$, or $\mu_i = t$. The equality case is a surrogate convention, not the operational tie-breaking rule.

For independent retrainings with marginal selection probabilities p_i and exact capacity, expected normalized overlap is $k^{-1}\sum_i p_i^2$. The surrogate approximates these population probabilities; it is distinct from the finite-S identity based on empirical selection frequencies.

Let τ be the mean exact-top-k cutoff and $p_i^G = p_i^G(\tau)$. The raw surrogate is

$$\widehat{\mathrm{AR}}_{\mathrm{raw}} = 1 - \frac{1}{k}\sum_i p_i^G(1 - p_i^G) = \frac{1}{k}\sum_i (p_i^G)^2 + r_{\mathrm{cap}}, \qquad r_{\mathrm{cap}} = 1 - \frac{1}{k}\sum_i p_i^G.$$

It replaces the run-specific cutoff by its mean, omitting cutoff variation and score–cutoff covariance. Unless $r_{\mathrm{cap}} = 0$, it also differs from the expected-intersection surrogate.

For capacity calibration, we numerically seek $\tau^\star$ such that $\sum_i p_i^G(\tau^\star) = k$. A unique solution exists when all $\sigma_i > 0$ and $0 < k < n$; zero-variance scores can introduce jumps. In all 35 evaluated cells, the numerical solution satisfies the capacity constraint to within 2.0×10^{-15}. Writing $p_i^\star = p_i^G(\tau^\star)$, calibration gives

$$\widehat{\mathrm{AR}}_{\mathrm{cap}} = 1 - \frac{1}{k}\sum_i p_i^\star(1 - p_i^\star) = \frac{1}{k}\sum_i (p_i^\star)^2.$$

Calibration constrains the marginal probabilities; it does not specify a joint model of exact top-k selection.

Across the 35 setting–budget cells, calibration raises the Spearman association with observed AR from 0.9706 to 0.9821 and reduces MAE from 0.0591 to 0.0427. Because the five budgets within each setting are dependent and the score statistics and observed AR come from the same retrainings, these comparisons are descriptive rather than out-of-sample prediction results.

C.9 COMMON-ERROR THRESHOLD SENSITIVITY

We evaluate sensitivity to the common-error threshold $q \in \{6, 8, 10\}$, corresponding to examples misclassified in at least 60%, 80%, or 100% of the ten retrainings. Table 12 reports the top-10% results. The qualitative cross-setting pattern is preserved across thresholds.

Table 12: Top-10% common-error diagnostics for $q \in \{6, 8, 10\}$.

Setting	CEFR 6	CEFR 8	CEFR 10	PairDis 6	PairDis 8	PairDis 10
R18/C10	0.8000	0.8199	0.7874	0.3171	0.3309	0.3216
R18/C100	0.5889	0.5875	0.5791	0.2329	0.2326	0.2296
R34/C10	0.8305	0.8397	0.8091	0.3141	0.3245	0.3170
R34/C100	0.6515	0.6523	0.6205	0.2686	0.2683	0.2549
D121/C10	0.7855	0.8238	0.8182	0.2943	0.3211	0.3264
R18/ImageNet-100	0.6040	0.6065	0.5867	0.2432	0.2439	0.2377
DBERT/Civil	0.2422	0.2648	0.3049	0.0876	0.0959	0.1114

C.10 ENTROPY VERSUS MSP-DERIVED UNCERTAINTY SELECTION

Table 13 compares the primary entropy ranking with $1 - \mathrm{MSP}$ at the top-10% budget.

Table 13: Entropy versus $1 - \mathrm{MSP}$ uncertainty selection at top-10%. Policy overlap is the mean normalized intersection $|\mathcal{S}_s^H \cap \mathcal{S}_s^M|/k$ across retrainings.

Setting	Ent. AR	MSP AR	Ent. CEFR	MSP CEFR	Ent. PairDis	MSP PairDis	Policy overlap
R18/C10	0.5981	0.5961	0.8199	0.8276	0.3309	0.3358	0.9838
R18/C100	0.5486	0.4832	0.5875	0.6893	0.2326	0.2782	0.8318
R34/C10	0.5849	0.5830	0.8397	0.8439	0.3245	0.3271	0.9876
R34/C100	0.4994	0.4221	0.6523	0.7468	0.2683	0.3149	0.8188
D121/C10	0.6238	0.6226	0.8238	0.8286	0.3211	0.3229	0.9889
R18/ImageNet-100	0.5454	0.4424	0.6065	0.7202	0.2439	0.3014	0.8040
DBERT/Civil	0.8868	0.8868	0.2648	0.2648	0.0959	0.0959	1.0000

C.11 TIE-BREAKING SENSITIVITY ON CIVIL COMMENTS

Civil Comments contains exact uncertainty-score ties near some operational cutoffs, but the tied blocks are small relative to review capacity. Across seeds, the largest cutoff tie block contains 24 examples, corresponding to 0.08% of the 30% review capacity. Table 14 compares the canonical ascending-index rule with descending-index and 20 fixed random orders. Across all five budgets, the maximum AR span is below 1.2×10^{-4}, indicating that the high Civil reproducibility is insensitive to the tested deterministic tie-breaking rules.

Table 14: Civil Comments tie-breaking sensitivity. Random-order results summarize 20 fixed sample permutations shared across all ten retrainings.

Budget	Canonical AR	Desc.-index AR	Random mean	Random min–max
1%	0.352065	0.351974	0.352015	0.351974–0.352065
5%	0.808248	0.808211	0.808182	0.808138–0.808257
10%	0.886838	0.886832	0.886834	0.886809–0.886857
20%	0.915460	0.915456	0.915459	0.915424–0.915479
30%	0.921759	0.921762	0.921754	0.921726–0.921765

C.12 PROPERTIES OF HIGH-INSTABILITY EXAMPLES

At the top-10% budget, we define the high-instability subset as examples selected by four to six of the ten retrainings, equivalently $\widehat{\mathrm{IDI}}_i \geq 0.96$. For each example, define pairwise prediction disagreement

as

$$Q_i = \binom{S}{2}^{-1} \sum_{s<t} \mathbb{I}[\hat{y}_{i,s} \neq \hat{y}_{i,t}], \tag{23}$$

and let AnyFlip indicate whether its predicted class changes across the ten runs. Table 15 summarizes prediction churn within this subset.

Table 15: Prediction churn within the top-10% high-instability subset.

Setting	N high-inst.	Pairwise Q	AnyFlip
R18/C10	547	0.206784	0.6344
R18/C100	640	0.558785	0.9453
R34/C10	577	0.209667	0.6308
R34/C100	733	0.603547	0.9782
D121/C10	470	0.163073	0.5383
R18/ImageNet-100	330	0.639300	0.9697
DBERT/Civil	1464	0.000683	0.0034

Stable class, changing allocation. Among examples whose predicted class is constant across all ten runs, the numbers that nevertheless change review status ($0 < m_i < 10$) are 1,156/8,903 (R18/C10), 331/6,247 (R18/C100), 1,226/8,873 (R34/C10), 308/6,247 (R34/C100), 1,222/9,016 (D121/C10), 101/2,846 (R18/ImageNet-100), and 5,933/94,875 (DBERT/Civil).

D Selective Utility and Chance Correction

Let $e_{i,s} = \mathbb{I}[\hat{y}_{i,s} \neq y_i]$. At capacity $k < n$, retained risk is

$$\frac{\sum_i (1 - Z_{i,s}) e_{i,s}}{n - k}, \tag{24}$$

and actual error coverage is

$$\frac{\sum_i Z_{i,s} e_{i,s}}{\sum_i e_{i,s}} \tag{25}$$

when errors exist. The latter uses each run's error set, whereas common-error coverage uses the fixed set $\mathcal{E}_q$.

AURC is

$$\frac{1}{n} \sum_{j=1}^{n} \frac{1}{j} \sum_{\ell=1}^{j} e_{i_{\ell,s},s}, \tag{26}$$

where retained examples are ordered by ascending entropy and, for ties, descending sample index, exactly reversing the review order. Table 16 reports across-run ranges of selective utility. Table 17 reports chance-corrected AR across review budgets. For exact capacity k, the chance baseline is k/n.

Relation to subset stability. Writing $a = k/n$ and $s_i^2 = S\hat{p}_i(1 - \hat{p}_i)/(S - 1)$, the equal-size subset identity gives

$$\widehat{\Phi} = 1 - \frac{n^{-1} \sum_i s_i^2}{a(1 - a)} = 1 - \frac{1 - \mathrm{AR}}{1 - a} = \mathrm{AR}_{\mathrm{adj}}. \tag{27}$$

Chance correction requires $0 < k < n$ and uses the actual integer capacity; for Civil at 1%, $k = 974$ and $n = 97{,}320$. It is a re-expression of the same overlap information rather than an independent confirmation.

Table 16: Maximum minus minimum across ten runs. Risk and actual error coverage use top-10% review; AURC uses the full risk–coverage curve. Values marked pp are percentage points.

Setting	Retained-risk range (pp)	Error-coverage range (pp)	AURC range
R18/C10	0.2667	4.2780	0.0019
R18/C100	0.5889	1.3446	0.0025
R34/C10	0.3222	5.9448	0.0016
R34/C100	1.3667	2.3929	0.0062
D121/C10	0.3556	6.5490	0.0023
R18/ImageNet-100	1.7111	2.6299	0.0058
DBERT/Civil	0.0548	1.1919	0.0001

Table 17: Chance-corrected AR across review budgets. This quantity equals the subset-stability estimator $\widehat{\Phi}$.

Setting	1%	5%	10%	20%	30%
R18/C10	0.2168	0.3859	0.5534	0.6280	0.5883
R18/C100	0.2121	0.3980	0.4984	0.5920	0.6263
R34/C10	0.2074	0.3655	0.5388	0.6104	0.5648
R34/C100	0.1540	0.3403	0.4437	0.5458	0.5940
D121/C10	0.2330	0.4024	0.5820	0.6046	0.5407
R18/ImageNet-100	0.2453	0.3924	0.4949	0.6027	0.6476
DBERT/Civil	0.3455	0.7982	0.8743	0.8943	0.8882