THE CURATION PARADOX: CLEANER LABELS CAN BE HARDER TO CORRECT

Anonymous authors
Paper under double-blind review

ABSTRACT

Cleaner labels and broader coverage are appealing goals for benchmark curation, but neither guarantees better model selection. We study speaker verification, where evaluation chooses both a model and an acceptance threshold. Agreement filtering can tie surviving label errors to model scores, violating the independence required by class-confusion correction. We express the remaining cost distortion as a covariance residual and test its consequences using frozen public speaker models with simulated label errors. An oracle intervention redistributes errors while preserving the cohort and error counts, isolating the role of error placement. Consistent identity merges and splits also produce cases where filtering improves labels but worsens the selected decision, alongside cases where filtering helps. For developers with only provisional labels, a separate budget study compares uniform, agreement, and score-stratified sampling. Broader observable coverage does not reliably improve decisions; closer agreement with the full provisional benchmark can preserve its errors. Moreover, selecting fewer pairs may save little encoder computation because pairs share recordings. These controlled results motivate evaluating curation by decision quality, information requirements, and processing cost, rather than treating label agreement or subset size as quality certificates.

1 INTRODUCTION

A benchmark can lose label errors and still become less useful for selecting a model. The location of the remaining mistakes matters: known noise rates correct population error curves under class-conditional independence, but errors tied to model decisions can invalidate that correction. We test whether retaining labels that agree with a curator creates this failure.

In speaker verification, the benchmark selects an encoder together with an acceptance threshold. Minimum detection cost (minDCF) expresses the deployment tradeoff between accepting different speakers and rejecting the same speaker. Noisy labels can distort that tradeoff; filtering also changes which trials enter the comparison. Label accuracy, score-curve accuracy, and the quality of the selected decision therefore require separate assessment.

Our test starts with established class-conditional noise correction. Agreement filtering ties the surviving mistakes to the curator's decisions, so even the exact class-confusion matrix may leave a distorted cost curve. We isolate this dependence by redistributing errors within each reference class while holding the retained examples, scores, and class-specific error counts fixed. Changes in decision quality then reflect error placement within the same cohort.

The experiments use five public checkpoints and two development/test studies with disjoint speakers. Redistributing errors reduces mean corrected retained-frame regret in all 24 conditions. A second experiment assigns consistent provisional identities to recordings through merges and splits: filtering again cleans labels, while decision effects remain mixed. These controlled tests establish a failure mode of curation, not a universal preference for noisier data.

Developers may have neither reference labels nor a trusted development set. We therefore also compare three subset designs restricted to provisional labels and at most one source model's scores. This study separates fidelity to the existing benchmark from reference decision quality and records the inference inputs each design requires. Its counterexamples expose two further limits: broader

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.

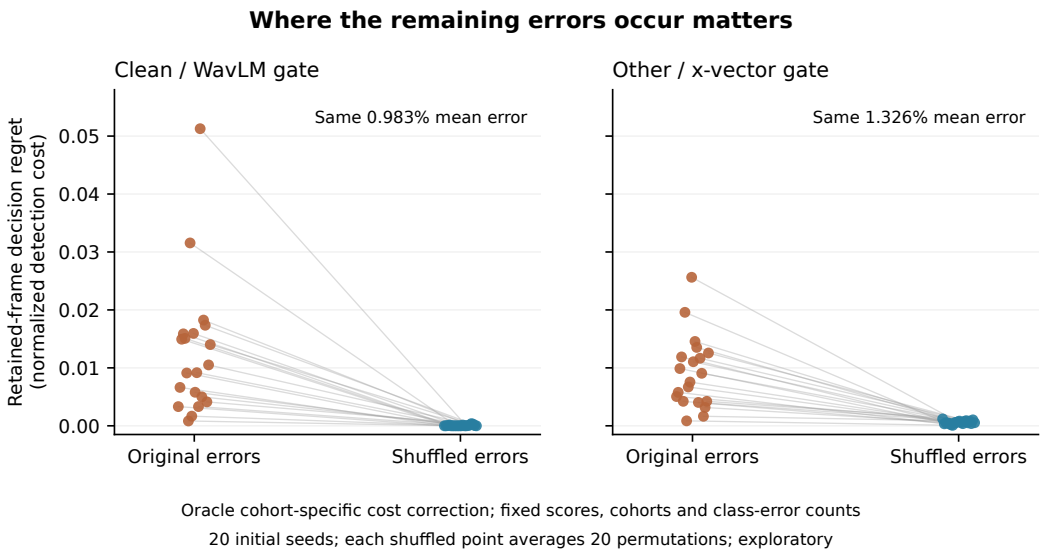

Figure 1: Error placement changes the consequence of the same label-error counts. A line connects an initial corruption draw to the mean of its 20 within-class permutations; there are 20 initial seeds. Each pair preserves the retained examples, scores, and class-specific error counts and uses identical oracle confusion correction. The outcome is retained-frame reference regret at prior 0.10. These exploratory contexts were chosen after observing adverse curation outcomes.

score coverage need not improve decisions, and a small pair budget need not substantially reduce encoder work. The contribution is an empirical and analytical account of these distinctions, rather than a new sampling algorithm or a way to recover truth without additional assumptions.

2 Related work

Models used to construct a benchmark can affect the relative performance measured on it (Phang et al., 2021). AFLite filters adversarially to reduce dataset artifacts (Le Bras et al., 2020), while SMART retains strong rank agreement despite removing many examples (Gupta et al., 2025). Arena-Hard likewise uses its BenchBuilder pipeline to construct evaluation data with models (Li et al., 2025). Together, these studies motivate testing the consequences of curation without presuming that filtering must harm comparisons.

Benchmark compression already uses historical model responses to approximate full evaluations, as in tinyBenchmarks and DISCO (Maia Polo et al., 2024; Rubinstein et al., 2026). Our budget study uses established score-stratified sampling (Bennett & Carvalho, 2010) to distinguish fidelity to a provisional benchmark from reference decision quality; it introduces no new subset-selection algorithm.

Reference-label mistakes can affect which model is selected (Northcutt et al., 2021), and revising reference annotations can change apparent progress (Beyer et al., 2020). We examine the additional decision made when a distorted score curve selects the operating threshold as well as the model.

Equal error rate (EER) locates the point where false-accept and false-reject rates coincide. For a specified deployment prior, minimum detection cost chooses the best threshold, whereas actual detection cost evaluates a given threshold (Sadjadi et al., 2021). Prior speaker work studies merged and split identities during training (Pham et al., 2020) and the sensitivity of bias conclusions to evaluation choices (Hutiri et al., 2024). Here, proxy labels enter evaluation after training and influence the selected system and threshold.

The theory of class-conditional noise already establishes correction, cost transformations, and conditions that preserve balanced-error or AUC orderings (Scott et al., 2013; Menon et al., 2015; Natarajan et al., 2018). Some structured instance-dependent noise also preserves classification or AUC

consistency (Menon et al., 2018); dependence on scores is therefore insufficient to establish harm. Earlier work identifies bias from conditional dependence (Vacek, 1985) and examines its effect on ROC correction with imperfect references (Sun et al., 2024). Discrepant analysis can favor an evaluated test by resolving only disagreements with an imperfect reference, even when the resolver is perfect (Miller, 1998). Its intervention changes reference assignments; ours changes retention and examines the resulting model–threshold choice. Training-time failures of oracle correction are also known (Feng et al., 2026). With encoders frozen, we test correction after agreement filtering by changing error placement while preserving the cohort and its error counts.

3 From label cleanliness to decision cost

Each trial compares an enrollment recording and a probe. Its reference label is $Y \in \{0, 1\}$, with one indicating a speaker match; $Z \in \{0, 1\}$ denotes the provisional label. At threshold t, model m makes the acceptance decision $h_m(t) = \mathbf{1}\{s_m \geq t\}$. The fixed gate g uses observable scores and proxy labels to retain cohort S_g. All quantities in a comparison use that same cohort and the same nonnegative trial weights, normalized to sum to one. Our target frames weight trials equally; probability-sampled subsets use the inclusion weights specified in Section 5.

Class-conditional costs. Within the fixed cohort, define $\mu_L = \mathbb{E}[L]$, $F_m(t) = \mathbb{E}[\mathbf{1}\{s_m < t\}]$, and $G^L_m(t) = \mathbb{E}[L\mathbf{1}\{s_m < t\}]$ for label source $L \in \{Y, Z\}$. Assume positive weighted mass in both classes, $0 < \mu_L < 1$, so that the conditional error rates are

$$\mathrm{FNR}^L_m(t) = \frac{G^L_m(t)}{\mu_L}, \qquad \mathrm{FPR}^L_m(t) = \frac{1 - F_m(t) - \mu_L + G^L_m(t)}{1 - \mu_L}. \tag{1}$$

With deployment prior $0 < p < 1$ and positive error costs $C_M, C_F > 0$, the normalized detection cost is

$$R^L_m(t) = \frac{C_M p\, \mathrm{FNR}^L_m(t) + C_F(1-p)\, \mathrm{FPR}^L_m(t)}{d_0}, \quad d_0 = \min\{C_M p, C_F(1-p)\}. \tag{2}$$

Trial-list prevalence and deployment prior are distinct quantities. Changing the labels can change both denominators in Equation 1; a valid DCF correction must account for these class masses as well as the ordinary-error difference.

3.1 Cleaning can change which errors are correctable

Noise prevalence and error placement affect correction differently. Suppose the pre-filter noise is independent of the full score vector conditional on the reference class: $Z \perp (s_1, \ldots, s_K) \mid Y$. Each proxy-class error curve then mixes the two reference-class curves. Established methods correct this mixture (Scott et al., 2013; Menon et al., 2015; Sun et al., 2024), providing a baseline against which to assess curation.

Let $a = \mathbb{P}(Y = 1 \mid Z = 1)$ and $b = \mathbb{P}(Y = 1 \mid Z = 0)$. With unit error costs and $0 < q < 1$, use the unnormalized risk $D^L_m(t; q) = q\, \mathrm{FNR}^L_m(t) + (1-q)\, \mathrm{FPR}^L_m(t)$ for $L \in \{Y, Z\}$. Include both constant policies, always-accept and always-reject, in the candidate set. The stated independence gives

$$D^Z_m(t; q) = k + A\, \mathrm{FNR}^Y_m(t) + B\, \mathrm{FPR}^Y_m(t), \tag{3}$$

with coefficients $A = qa - (1-q)b$, $B = (1-q)(1-b) - q(1-a)$, and $k = q(1-a) + (1-q)b$. When $A, B > 0$, minimizing proxy cost is equivalent to minimizing reference cost at effective prior $A/(A+B)$. When $a > b$ and $q \leq b/(a+b)$, even a perfect scorer has an optimal always-reject policy. At our test prevalence $1/29$ with 10% symmetric noise, the boundary is 0.01599, above the nominal prior 0.01. Thus population-level cost distortion alone can produce rejection collapse on the unfiltered benchmark; finite corruption draws need not attain that population optimum.

To target reference prior p, apply the established cost correction at

$$q^*(p) = \frac{b + p(1-2b)}{a + b - 2p(a+b-1)}, \tag{4}$$

under $a > b$, $0 < p < 1$, and positive weighted class masses for both label sources. The correction leaves labels intact and adjusts the selection objective. With q^*, Equation 3 becomes a positive affine transform of the target reference risk; the mixture model therefore gives identical model–threshold optima under the two objectives.

Agreement retention can break the required independence because $S_g = \{Z = c\}$ forces each retained proxy to equal the curator decision. Within any retained reference class where c varies, that decision determines which labels are wrong. A fixed gate based only on scores and independent randomization preserves the joint independence when retained class masses are positive. An agreement gate also conditions on the proxy label.

To characterize the remaining distortion, fix a labeled frame with $a > b$ and positive class masses under each label source. Its realized confusion proportions determine $q_p = q^*(p)$. Define $F = \mathbf{1}\{Z \neq Y\}$, $\mu = \mathbb{P}(Y = 1)$, and the residual

$$K_m(t) = \mu \operatorname{Cov}(F, h_m(t) \mid Y = 1) - (1 - \mu) \operatorname{Cov}(F, h_m(t) \mid Y = 0). \tag{5}$$

The following equality holds even when errors depend on scores:

$$D_m^Z(t; q_p) = C_{q_p} + \lambda_p D_m^Y(t; p) + \gamma_{q_p} K_m(t), \tag{6}$$

Here C_{q_p} is constant across policies, while $\lambda_p, \gamma_{q_p} > 0$ are determined by the class-confusion quantities and prior; Appendix A derives them. Exact confusion counts therefore leave a score-dependent term that can change with the threshold. For agreement-retained trials, $K_m(t)$ equals the negative class-weighted within-class covariance between curator and evaluator decisions. A positive weighted covariance reduces corrected proxy cost relative to its affine reference component.

Uniformly permuting proxy labels within each reference class of a uniformly weighted frame preserves its examples and class-specific error counts, with $\mathbb{E}[K_m(t)] = 0$ for every fixed policy. A realized permutation can retain residual structure that optimization exploits. We measure the resulting decision regret because pointwise mean correction does not determine the outcome of finite-sample policy selection.

3.2 Measure the cost of the selected decision

Write $(\hat{m}, \hat{t})$ for the policy chosen by the benchmark objective after any stated noise correction. For each model, let t_m^Y minimize reference cost over its frozen threshold policies. The threshold contribution is $\Gamma_{\hat{m}} = R_{\hat{m}}^Y(\hat{t}) - R_{\hat{m}}^Y(t_{\hat{m}}^Y)$, giving the total reference regret

$$R_{\hat{m}}^Y(\hat{t}) - \min_{m,t} R_m^Y(t) = \Gamma_{\hat{m}} + \left[\min_t R_{\hat{m}}^Y(t) - \min_{m,t} R_m^Y(t) \right]. \tag{7}$$

The first term measures loss from the threshold choice; the second measures loss from the model choice. Both are nonnegative. Reordering nearly tied models may have little decision cost, whereas preserving their order can still leave a poor threshold. We report decision regret, with threshold conventions and population effects specified in Appendix B.

The selection population. Filtering changes the trial population used for selection. Our primary outcome evaluates the selected policy on the complete original reference frame. The independent-pair study additionally reports retained-cohort outcomes, because a good choice on that cohort can perform poorly on the original frame. Interpreting exclusion as an operational rejection would additionally require accounting for its downstream cost.

4 Controlled experiments with real speaker models

Speaker frames and model roles. We evaluate SpeechBrain ECAPA-TDNN and x-vector, WavLM Base-Plus for speaker verification, and WeSpeaker ResNet34-LM and ResNet221-LM (Desplanques et al., 2020; Snyder et al., 2018; Chen et al., 2022; Wang et al., 2023). Two studies use the clean and other acoustic splits of LibriSpeech, an English read-speech corpus (Panayotov et al., 2015). Each eligible speaker supplies four enrollment and four probe recordings from distinct chapters; each input is at most six seconds long. The clean study uses 31 development and 29 test

speakers, and the other study uses 28 and 29. Each test frame comprises 13,456 pairs, of which 464 are same-speaker trials. All four speaker sets are disjoint. Reference labels come from dataset speaker IDs, without new human identity adjudication. Table 1 reports reference performance on these frames.

Table 1: Reference performance on the complete test frames, with each input limited to six seconds. EER is a percentage; normalized minDCF uses priors 0.01 and 0.10 and unit costs. The chapter-disjoint frames are constructed for this study and are distinct from standard leaderboard trial lists.

Model	Clean EER	$C_{.01}$	$C_{.10}$	Other EER	$C_{.01}$	$C_{.10}$
ECAPA	0.49	0.062	0.027	1.49	0.118	0.048
x-vector	5.93	0.418	0.221	9.27	0.570	0.354
WavLM	5.17	0.557	0.286	6.98	0.681	0.365
ResNet34	0.34	0.050	0.021	0.85	0.117	0.042
ResNet221	0.22	0.037	0.013	0.65	0.102	0.033

Corruption, curation, and decision outcomes. We independently flip pair labels with probability 0.10 or 0.20 across 20 fixed seeds. ECAPA, x-vector, and WavLM each serve as curator, using the deterministic threshold nearest their own development equal-error point. Evaluation uses either the full frame or the pairs on which the curator agrees with the noisy proxy. Each primary model-choice set excludes its curator and contains four evaluators. An uncorrupted-label control separates population selection from label corruption. Target speaker references are never used to train or fine-tune an encoder.

Our primary outcome is the original-frame reference regret of the model and threshold chosen by proxy minDCF: its reference cost minus the eligible reference oracle's cost on that original frame. This counterfactual comparison conditions on a finite frame; it does not measure performance on future traffic. Supplementary outcomes include EER, retained-frame regret components, strict wrong winners, and all proxy-optimal tie ranges. We recompute class denominators and metric thresholds for each label source. Both costs equal one; target priors are 0.01 and 0.10. Appendix C records the protocol, model versions, and panel expansion; Appendix D provides the full grid.

Compare corrections under explicit information assumptions. Equation 4 gives three correction baselines for the frozen scores and corruption draws. Nominal correction uses the imposed flip rate and exact target reference prevalence; realized correction uses the observed target confusion counts. Both are oracle baselines. Transferred correction estimates forward flip rates from all trusted development labels under independent corruption, then infers target prevalence from proxy labels. It requires trusted development references and assumes the noise mechanism transfers. Table 4 keeps these information regimes separate.

The central comparison corrects the unfiltered and retained frames using each frame's *own realized confusion matrix*. Unfiltered correction yields lower mean original-frame regret in every clean-speech condition and half of the other-speech conditions. Of the pointwise paired intervals across the complete grid, 15 favor unfiltered correction and two favor corrected gating. Cleaner labels therefore do not guarantee a better corrected decision; the direction also varies across populations. Because this comparison changes the selected population, we next hold the cohort fixed.

Change error placement within a fixed cohort. For each initial corruption draw, we freeze the agreement-retained cohort and permute flip indicators within each true class, using 20 permutations per initial seed. Every score, retained example, and class-specific error count stays fixed. Both error placements use the same oracle cohort correction. Figure 1 shows two previously selected contexts with 10% noise and prior 0.10; Appendix A.1 includes every condition. Permutation lowers mean retained-frame regret for every initial seed in the two illustrated contexts and lowers all 24 condition means in the full extension. This direction is consistent with eliminating the covariance term in expectation. The observed outcome, however, includes finite-sample optimization and cannot be inferred by exchanging expectation and minimization. The full-grid extension followed the two selected examples and is exploratory. Because it uses reference error locations, it diagnoses the mechanism rather than providing a deployable repair.

Measure consequences on the original frame. We evaluate each saved model–threshold policy on the complete original frame without reselecting either component. Mean original-frame regret falls in 23 of 24 conditions; 13 comparisons pass Holm adjustment across the 24 added contrasts: ten at prior 0.10 and three at prior 0.01. The exception is other speech with the x-vector curator, 10% noise, and prior 0.01: mean regret rises, but its interval includes zero. For the illustrated clean/WavLM and other/x-vector contexts, regret falls from 0.03160 to 0.00223 and from 0.03067 to 0.00664, respectively. Appendix A.1 retains every condition. Retained-frame regret isolates error placement within a fixed cohort; original-frame regret also captures the consequence of selecting a policy on a different population. Both outcomes condition on the observed frames rather than future traffic.

Benchmark against trusted development selection. Selecting the model and threshold on trusted development data alone gives the lowest mean original-frame regret among the displayed options without target-oracle information in 20 of 24 conditions (Table 4). This comparator requires existing trusted labels and is unavailable in the strict provisional-only setting. Without correction, agreement filtering improves mean regret in 19 of 24 conditions, including every condition at prior 0.01 (Appendix D). These outcomes support a conditional failure of correction, not a general recommendation against filtering.

4.1 Does the failure survive consistent identity labels?

Speaker-label merges and splits are established corruption operators (Pham et al., 2020); here we apply them to frozen evaluator–threshold selection rather than encoder training. Each recording receives one provisional identity, reused in every trial. Across both acoustic panels, low/high severity combines two/five disjoint speaker-pair merges with two/five balanced recording splits. Merges are either random or chosen using the same fixed ECAPA similarity rule for every active curator; the latter introduces score dependence before filtering. Both mechanisms are simulated and use reference speaker groups to construct the errors. Eight frame/mechanism/severity settings, three trusted-development curator gates, and two priors give 48 regret conditions, each averaged over 20 new seeds. Each comparison excludes the active curator from model selection and the reference oracle. Table 2 includes every frame, mechanism, and severity.

Agreement filtering lowers mean label error in all 24 distinct frame/mechanism/severity/curator conditions, retaining 88.19–98.48% of trials on average. Initial mean error ranges from 0.606% to 1.535%, versus 0.002% to 0.924% after filtering. Nevertheless, mean original-frame regret rises in 32 of 48 unadjusted comparisons and 33 of 48 comparisons using each frame's own oracle confusion correction; the other 16 and 15 comparisons improve. Enumerating all proxy-optimal policies on the original score grid leaves 28 higher-regret mean comparisons for each method even under the most favorable tie choices for retention. Exact-count pair controls are also mixed, with 29 and 34 higher-regret means. Thus identity consistency does not remove the counterexample, but these results neither establish universal filtering harm nor attribute differences from the earlier 10%/20% corruption study solely to identity structure. All uncertainty summaries condition on the fixed frames and declared simulation; near-merge pairs are fixed across seeds.

For scale, two post hoc illustrations use the same clean, similarity-based high-severity partitions at prior 0.10. WavLM retention lowers label error from 1.525% to 0.924% while corrected regret rises from 0.07120 to 0.21382; ECAPA retention instead lowers regret from 0.15979 to 0.02063. Each gate has its own curator-excluded oracle. Appendix E specifies the partitions, matched controls, and complete identity-grid means.

5 What a budgeted provisional benchmark preserves

A developer without trusted labels can test fidelity to the existing provisional benchmark, but cannot thereby establish reference quality. We distinguish the two by evaluating a subset-selected policy against the full-frame optimum under Z and, separately, under withheld Y, using the same four eligible models and operating prior. These regrets are denoted G_Z and G_Y. Only the offline research harness sees the reference identities and full candidate scores.

Table 2: Identity-consistent corruption and exact-count pair controls. Low/high severity applies two/five disjoint speaker merges and two/five balanced recording splits. Error is the initial identity-label error percentage averaged over 20 seeds; the pair control matches each seed's false-negative and false-positive counts exactly. Entries count conditions where agreement filtering increases mean original-frame regret, out of six curator/prior combinations; all remaining conditions improve. "Corrected" gives each frame its own oracle confusion matrix. These dependent, exploratory finite-frame counts are descriptive, not significance tests.

				Identity partition		Matched pairs	
Frame	Merges	Level	Error %	Raw	Corrected	Raw	Corrected
Clean	Random	Low	0.609	5/6	5/6	5/6	5/6
Clean	Random	High	1.535	3/6	5/6	3/6	6/6
Clean	ECAPA-near	Low	0.607	6/6	6/6	6/6	6/6
Clean	ECAPA-near	High	1.525	5/6	4/6	3/6	5/6
Other	Random	Low	0.607	3/6	4/6	3/6	3/6
Other	Random	High	1.520	3/6	2/6	3/6	3/6
Other	ECAPA-near	Low	0.606	3/6	4/6	3/6	3/6
Other	ECAPA-near	High	1.524	4/6	3/6	3/6	3/6

A restricted, implementable comparison. Using the original independent-pair corruption draws, we compare uniform sampling, sampling within a curator-agreement gate, and probability sampling across curator-score octiles crossed with Z. The last design uses standard score stratification and inverse inclusion weights (Bennett & Carvalho, 2010; Horvitz & Thompson, 1952); it is a coverage baseline. Every selector receives existing provisional labels and at most one curator score per trial. Four held-out evaluators are revealed only on the chosen rows. The agreement gate uses a provisional-label EER threshold, unlike the trusted-development gates in Section 4. No trusted labels or noise rates enter selection or fitting.

Budgets are 128, 512, and 2048 pairs. Each nonempty coverage stratum receives samples with positive, known inclusion probability; weighted class totals then equal their known full-Z totals. Thus its DCF is design-unbiased for a fixed policy, although the optimized minimum and chosen policy are not generally unbiased. Uniform sampling uses estimated class totals. Agreement sampling gives excluded trials zero probability and estimates a different, retained-population objective. Appendix E fixes allocation and threshold conventions.

Table 3: Budgeted provisional evaluation. G_Z measures full-provisional decision regret; G_Y measures reference regret of the same policy. Means equally weight two frames, three excluded curators, two corruption rates, and 20 seed/draw pairs at each prior. Passes count encoder–recording evaluations, including the source pass for score-aware methods; they do not weight architectures by runtime. The full-provisional row is the offline validation limit.

		$p = 0.01$		$p = 0.10$		
Rows	Selection	G_Z	G_Y	G_Z	G_Y	Passes
128	Uniform	1.0521	0.9772	0.0886	0.2431	622.2
128	Agreement	8.7055	8.3780	0.8611	0.9539	849.9
128	Coverage	1.5322	1.3437	0.1136	0.2437	857.0
512	Uniform	0.2429	0.4964	0.0229	0.1546	917.9
512	Agreement	4.8670	4.4525	0.6482	0.7122	1146.8
512	Coverage	0.2715	0.4576	0.0279	0.1453	1149.0
2048	Uniform	0.0542	0.7077	0.0064	0.0897	928.0
2048	Agreement	2.7551	2.3037	0.5487	0.6038	1160.0
2048	Coverage	0.0792	0.6517	0.0076	0.0794	1160.0
13,456	Full provisional	0	0.8472	0	0.0350	928

More coverage does not establish better decisions. At budget 128, coverage sampling includes every score/label cell, versus 95.2% on average for uniform sampling, but has higher mean provisional and reference regret. Uniform sampling gives lower mean provisional regret than coverage

sampling at every budget and prior in Table 3, with no source-model pass. Coverage has modestly lower mean reference regret at the two larger budgets; unknown references cannot justify choosing it on that basis. The cells describe one curator's scores and Z, not a complete taxonomy of acoustic or demographic conditions.

Better fidelity can preserve a worse decision. At prior 0.01, increasing uniform sampling from 512 to 2048 pairs lowers mean provisional regret from 0.2429 to 0.0542 but raises reference regret from 0.4964 to 0.7077, toward the full-provisional policy's 0.8472. At prior 0.10 both regrets improve. This is a failure of provisional fidelity as a quality certificate, not evidence that smaller samples are generally preferable. A policy learned on a subset can have normalized regret above one when applied to the original population.

Count recordings as well as pairs. At 512 pairs, only 3.8% of the trial list is selected, yet uniform sampling touches an average 229.5 of 232 recordings. Four evaluated encoders therefore require 917.9 recording passes, compared with 928 for the full frame; coverage additionally needs a curator pass, totaling 1149.0. Cached scores simulate these access requirements. Pair accesses, recording passes, and architecture-dependent inference time are distinct costs: this experiment establishes neither a wall-clock speedup nor an advantage at matched total compute.

6 Discussion and limitations

The experiments distinguish three outcomes that need not improve together: label accuracy, the decision supported by a benchmark, and the computation needed to evaluate it. Agreement retention can improve the first while worsening the second. Class-confusion correction can fail to resolve this tradeoff when the remaining errors depend on scores, and identity consistency alone does not prevent it.

Practical takeaway. Specify the model–threshold decision, operating prior, and population before selecting a subset. To approximate the existing provisional benchmark, retain nonzero inclusion probabilities across the intended population and use the corresponding sampling weights; deterministic exclusions change that target. Report observable score/label coverage and provisional decision fidelity separately from reference accuracy. Charge the source model and unique recordings processed by every evaluator. If trusted development labels already exist, include development-only selection as a baseline; these experiments do not justify acquiring or assuming such labels when they are unavailable.

The evidence supports a conditional failure mode, not a universal preference for unfiltered data. Some curators and operating points benefit from retention, with and without correction. A ranking alone can also miss a costly threshold change. Evaluating the selected model and threshold on the specified original frame makes these distinctions visible.

All studied corruption is imposed relative to dataset speaker IDs on two small English read-speech frames. The merge/split experiment uses consistent provisional identity assignments but does not measure natural metadata errors. Recorded uncertainty conditions on the fixed speakers, scores, and declared randomization. Shared recordings and related checkpoints preclude treating trials, models, or grid cells as independent population replications. The budget study uses cached scores, measures a narrow score-based notion of coverage, and does not establish a wall-clock speedup or an advantage at matched total compute. Transfer to new speakers, languages, channels, and naturally corrupted labels remains untested.

7 Conclusion

A cleaner benchmark need not support a better model–threshold decision. The covariance residual explains why exact class-confusion correction can fail after agreement retention; controlled error placement and globally consistent identity corruption test the consequence. For provisional-only evaluation, subset fidelity, reference quality, and processing cost must be assessed separately. None can be inferred from label agreement or subset size alone.

AI USE STATEMENT

Generative AI supported idea development, literature searches, review of code and mathematics, and drafting. Executable analyses generated the empirical results, which were reconciled with saved outputs; source checks verified the bibliographic claims.

ETHICS STATEMENT

The experiments use released LibriSpeech audio and dataset speaker IDs. Imposed label corruption simulates metadata errors; it makes no claim about recording authenticity or contributor intent. No new human judgments are collected, and the reproduction package contains no private recordings or account information.

REPRODUCIBILITY STATEMENT

The retained reproduction record includes the protocol, model revisions and hashes, utterance manifests, inference code, score matrices, and analysis outputs. Selection protocols identify their permitted inputs, seeds, and inclusion probabilities. These records document executed exploratory studies; they do not constitute external preregistration. The portable package contains no private source material.

REFERENCES

Paul N. Bennett and Vitor R. Carvalho. Online stratified sampling: Evaluating classifiers at web-scale. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 1581–1584. ACM, 2010. doi: 10.1145/1871437.1871677. URL https://www.cs.cmu.edu/~vitor/papers/cikm2010-carvalho.pdf.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020. URL https://arxiv.org/abs/2006.07159.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113. URL https://arxiv.org/abs/2110.13900.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650. URL https://www.isca-archive.org/interspeech_2020/desplanques20_interspeech.html.

Chen Feng, Zhuo Zhi, Zhao Huang, Jiawei Ge, Ling Xiao, Nicu Sebe, Georgios Tzimiropoulos, and Ioannis Patras. Deconstructing the failure of ideal noise correction: A three-pillar diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34512–34523, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Feng_Deconstructing_the_Failure_of_Ideal_Noise_Correction_A_Three-Pillar_Diagnosis_CVPR_2026_paper.pdf.

Vipul Gupta, Candace Ross, David Pantoja, Rebecca J. Passonneau, Megan Ung, and Adina Williams. Improving model evaluation using SMART filtering of benchmark datasets. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4595–4615. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.235. URL https://aclanthology.org/2025.naacl-long.235/.

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

10.1080/01621459.1952.10483446. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483446.

Wiebke Hutiri, Tanvina Patel, Aaron Yi Ding, and Odette Scharenborg. As biased as you measure: Methodological pitfalls of bias evaluations in speaker verification research. In *Interspeech 2024*, pp. 4268–4272, 2024. doi: 10.21437/Interspeech.2024-1158. URL https://www.isca-archive.org/interspeech_2024/hutiri24_interspeech.html.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux. Libri-Light: A benchmark for ASR with limited or no supervision. *arXiv preprint arXiv:1912.07875*, 2019. URL https://arxiv.org/abs/1912.07875.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1078–1088. PMLR, 2020. URL https://proceedings.mlr.press/v119/bras20a.html.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-Hard and Benchbuilder pipeline. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34209–34231. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25h.html.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Aditya Krishna Menon, Brendan van Rooyen, Cheng Soon Ong, and Robert C. Williamson. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 125–134. PMLR, 2015. URL https://proceedings.mlr.press/v37/menon15.html.

Aditya Krishna Menon, Brendan van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107:1561–1595, 2018. doi: 10.1007/s10994-018-5715-3. URL https://link.springer.com/article/10.1007/s10994-018-5715-3.

W. C. Miller. Bias in discrepant analysis: When two wrongs don't make a right. *Journal of Clinical Epidemiology*, 51(3):219–231, 1998. doi: 10.1016/S0895-4356(97)00264-3. URL https://pubmed.ncbi.nlm.nih.gov/9495687/.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Cost-sensitive learning with noisy labels. *Journal of Machine Learning Research*, 18(155):1–33, 2018. URL https://www.jmlr.org/papers/v18/15-226.html.

Curtis Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/f2217062e9a397a1dca429e7d70bc6ca-Abstract-round1.html.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://ieeexplore.ieee.org/document/7178964.

Minh Pham, Zeqian Li, and Jacob Whitehill. How does label noise affect the quality of speaker embeddings? In *Interspeech 2020*, pp. 3216–3220, 2020. doi: 10.21437/Interspeech.2020-1395. URL https://www.isca-archive.org/interspeech_2020/pham20b_interspeech.html.

Jason Phang, Angelica Chen, William Huang, and Samuel R. Bowman. Adversarially constructed evaluation sets are more challenging, but may not be fair. *arXiv preprint arXiv:2111.08181*, 2021. URL https://arxiv.org/abs/2111.08181.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021. URL https://arxiv.org/abs/2106.04624.

Alexander Rubinstein, Benjamin Raible, Martin Gubri, and Seong Joon Oh. DISCO: Diversifying sample condensation for efficient model evaluation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/19c145aaad40927c51f4d10eaa339c20-Abstract-Conference.html.

Omid Sadjadi, Craig Greenberg, Elliot Singer, Lisa Mason, and Douglas Reynolds. NIST 2021 speaker recognition evaluation plan. Technical report, National Institute of Standards and Technology, 2021. URL https://www.nist.gov/system/files/documents/2021/07/12/2021_SRE_Evaluation_Plan_V5.pdf.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 489–511. PMLR, 2013. URL https://proceedings.mlr.press/v30/Scott13.html.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-Vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5329–5333. IEEE, 2018. doi: 10.1109/ICASSP.2018.8461375. URL https://www.danielpovey.com/files/2018_icassp_xvectors.pdf.

Jiarui Sun, Chao Tang, Wuxiang Xie, and Xiao-Hua Zhou. Nonparametric receiver operating characteristic curve analysis with an imperfect gold standard. *Biometrics*, 80(3):ujae063, 2024. doi: 10.1093/biomtc/ujae063. URL https://academic.oup.com/biometrics/article/80/3/ujae063/7712597.

P. M. Vacek. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41(4):959–968, 1985. URL https://pubmed.ncbi.nlm.nih.gov/3830260/.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. WeSpeaker: A research and production oriented speaker embedding learning toolkit. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2023. URL https://arxiv.org/abs/2210.17016.

A NOISE CORRECTION AND THE ERROR-PLACEMENT INTERVENTION

The derivation uses unnormalized detection risk D with unit error costs. Normalization preserves the minimizers at a fixed prior, but changing the proxy prior changes the tradeoff and makes numerical costs unsuitable for direct comparison. Every reported decision regret is evaluated in reference DCF at the desired prior p.

Use the equivalent notation $D_L^q(m,t) = D_m^L(t;q)$ for $L \in \{Y, Z\}$. Throughout, require $0 < p < 1$, positive mass for each reference and proxy class, and $a > b$. Set $q_p = q^*(p)$ from the fixed frame's realized confusion proportions and define $\nu = \mathbb{P}(Z = 1)$. Retaining the definitions of F, μ, and $K_m(t)$ in Equation 5, Equation 6 has coefficients

$$C_q = q(1-a) + (1-q)b, \qquad \lambda_p = \frac{a-b}{a+b+2p(1-a-b)}, \qquad \gamma_q = \frac{q}{\nu} + \frac{1-q}{1-\nu}.$$

On an agreement-retained cohort, $Z = c$ reduces the residual to $K_m(t) = -\mu \operatorname{Cov}(c, h_m(t) \mid Y = 1) - (1-\mu)\operatorname{Cov}(c, h_m(t) \mid Y = 0)$. The confusion correction therefore retains a threshold-dependent component: positive class-weighted covariance with the curator lowers the proxy cost relative to its affine reference component. The shuffle control below separates this residual structure from changes in the cohort or its error counts, while preserving the distinction between a mean cost curve and its optimized decisions.

Proof of the exact residual identity. On the fixed frame, define $\rho_y = \mathbb{P}(F = 1 \mid Y = y)$, $u = \mathrm{FNR}_Y$, and $v = \mathrm{FPR}_Y$. Decomposing each conditional product into its mean product and covariance gives

$$\mathbb{P}(Z = 1, h = 0) = \mu(1-\rho_1)u + (1-\mu)\rho_0(1-v) + K, \tag{8}$$

$$\mathbb{P}(Z = 0, h = 1) = \mu\rho_1(1-u) + (1-\mu)(1-\rho_0)v + K. \tag{9}$$

Dividing by ν and $1-\nu$ converts these joint masses to conditional error rates. Weighting the rates by their operating costs assigns coefficient γ_q to K; the remaining terms recover Equation 3, with the additional term $\gamma_q K$ accounting for dependence. Substituting Equation 4 yields $q_p a - (1-q_p)b = \lambda_p p$ and $(1-q_p)(1-b) - q_p(1-a) = \lambda_p(1-p)$ and establishes Equation 6. The condition $a > b$ ensures $\lambda_p > 0$. Let $\hat{\theta}$ minimize corrected proxy cost and θ^* minimize reference cost over the same model–threshold class. Their optimality implies

$$0 \le D_Y^p(\hat{\theta}) - D_Y^p(\theta^*) \le \frac{\gamma_{q_p}}{\lambda_p}[K(\theta^*) - K(\hat{\theta})] \le \frac{\gamma_{q_p}}{\lambda_p} \operatorname{osc}_\theta K(\theta). \tag{10}$$

Dividing by $\min(p, 1-p)$ converts the bound to normalized reference DCF. Agreement retention gives $F = 1 - c$ in class $Y = 1$ and $F = c$ in class $Y = 0$, establishing the curator-covariance formula. Every identity compares risks on one cohort; filtering's change of population remains a separate effect.

Scope of the shuffle control. The intervention uses uniformly weighted frames and a uniform permutation of Z within each reference class. This fixes a, b, μ, ν and gives $\mathbb{E}[K_m(t)] = 0$ at each fixed decision vector, so the mean cost curve satisfies the affine correction before optimization. Any individual permutation can leave nonzero K that influences the selected policy. The comparison therefore measures optimized regret against its permutation distribution; the pointwise expectation alone supplies no equality of optimized outcomes.

Exploratory controls. The controls were designed after inspecting the original results and remain exploratory. We first selected two adverse contexts for permutation, then expanded to the full 24-cell corpus/noise/curator/prior grid using the original 20 seeds and score matrices. Nominal-oracle correction combines the imposed transition rate with true frame prevalence; realized-oracle correction instead uses the frame's observed confusion counts. The direct corrected-versus-corrected comparison supplies each retained cohort with its own counts. Reusing original-frame counts after gating is an intentionally misspecified diagnostic.

Both transferred variants use an independent development corruption draw generated with the original seed plus 1,000,000. The first directly transfers posterior confusion proportions from development data. The second estimates development forward flip rates and infers target prevalence from

the observed proxy-positive fraction. This inference requires a transferable forward noise mechanism and all development reference labels, which lie outside a limited acquisition budget. None of the inferred prevalences required clipping. A separate expected-count control samples a fixed-frame trial uniformly and then draws its noisy label; it recovers the reference oracle in every tested condition. Its ratios of expected counts must be distinguished from expected ratios in finite corrupted frames.

Each initial corruption seed fixes a retained cohort and supplies 20 permutations within true classes, with no subsequent gating. Uncertainty is summarized across initial seeds, pairing each original outcome with its mean permuted outcome. The 400 permutations per context therefore do not constitute independent speaker replications. We use pointwise paired t intervals for the initial two contexts and additionally report Holm adjustments for the families defined in the complete-grid tables. Matching the confusion matrix requires oracle information; the experiment diagnoses error dependence without supplying a way to locate or repair unknown mistakes.

Which correction and comparator? The direction counts depend on the comparator. Relative to an unadjusted gate, unfiltered nominal-oracle correction has lower mean regret in 18 of 24 conditions (12 of 12 clean, 6 of 12 other); the realized-oracle version does so in 19 (12 clean, 7 other), and the transferred-rate version in 18 (11 clean, 7 other). Against a gate corrected with its own oracle counts, realized unfiltered correction has lower mean regret in 18 conditions (12 clean, 6 other). These counts are computed before rounding. Table 4 also includes the development-only policy, which selects its encoder and threshold without target references and excludes the active curator.

Residual diagnostic. The oracle diagnostic in Figure 2 connects the cost identity to the observed decisions: permutation lowers mean residual oscillation in all 12 distinct frame/curator/noise conditions. Regret also depends on the operating prior and gaps between competing policy costs. The worst-case bound is loose, and this same-frame comparison does not establish predictive accuracy.

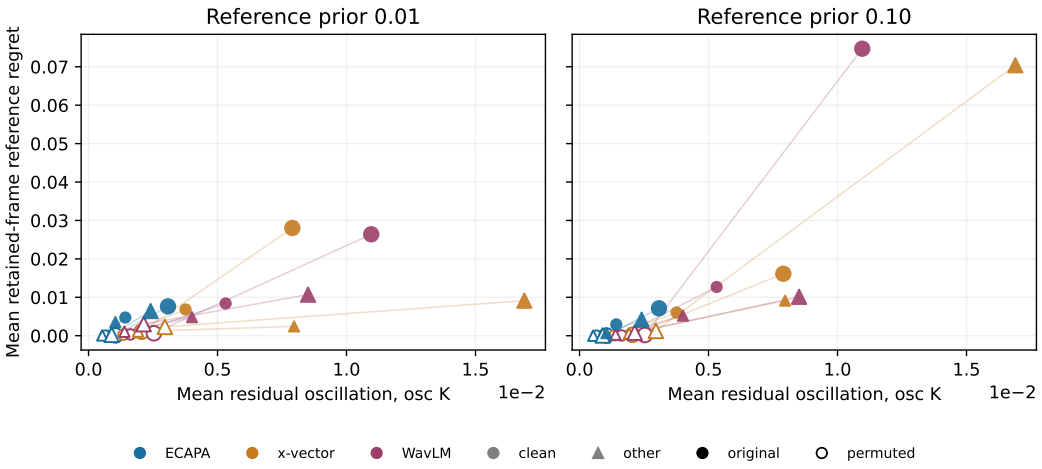

Figure 2: Oracle residual oscillation versus retained-frame decision regret. Each line joins the original error placement (filled) to its within-class permutations (open), with the cohort and class-specific error counts fixed. Points average 20 initial seeds; permuted points also average 20 shuffles per seed. Color denotes curator, shape denotes acoustic split, and larger markers indicate higher imposed noise. Oscillation spans every eligible model–threshold policy, excluding the curator. Panels use different reference priors, and both axes require reference labels.

Table 4: Original-frame normalized reference regret over 20 corruption seeds. Unfiltered nominal correction uses the imposed noise rate and oracle target prevalence; realized correction uses the full frame's oracle confusion counts. Transferred correction estimates forward noise rates from all trusted development references and target prevalence from proxy labels, assuming rate transfer. Retained oracle correction uses the retained cohort's own confusion counts. Dev-only selects model and threshold using trusted development references alone. Every selection excludes the curator. Development labels and model scores are outside the simulated acquisition budget. These exploratory means are dependent; favorable-condition counts use full precision rather than the five-decimal display.

				Unfiltered			Retained		Dev-only
Frame	Flip %	p	Curator	Nominal	Realized	Transferred	Proxy	Oracle	Policy
Clean	10	0.01	ECAPA	0.00741	0.00859	0.00584	0.02458	0.02458	0.02155
Clean	10	0.01	WavLM	0.01128	0.00876	0.01315	0.01907	0.04044	0.02155
Clean	10	0.01	x-vector	0.01128	0.00876	0.01315	0.03646	0.03907	0.02155
Clean	10	0.10	ECAPA	0.00226	0.00226	0.00234	0.00744	0.00854	0.00231
Clean	10	0.10	WavLM	0.00226	0.00226	0.00234	0.02057	0.03160	0.00231
Clean	10	0.10	x-vector	0.00226	0.00226	0.00234	0.00864	0.01123	0.00231
Clean	20	0.01	ECAPA	0.02447	0.02415	0.02932	0.02469	0.02704	0.02155
Clean	20	0.01	WavLM	0.02824	0.02791	0.02824	0.04407	0.06552	0.02155
Clean	20	0.01	x-vector	0.02824	0.02791	0.02824	0.04351	0.05718	0.02155
Clean	20	0.10	ECAPA	0.00533	0.00413	0.00531	0.01684	0.01750	0.00231
Clean	20	0.10	WavLM	0.00576	0.00456	0.00575	0.06043	0.13822	0.00231
Clean	20	0.10	x-vector	0.00576	0.00456	0.00575	0.01409	0.02355	0.00231
Other	10	0.01	ECAPA	0.00728	0.00735	0.00607	0.01039	0.01039	0.00115
Other	10	0.01	WavLM	0.00994	0.00973	0.00765	0.00747	0.00478	0.00115
Other	10	0.01	x-vector	0.00994	0.00973	0.00765	0.00759	0.00568	0.00115
Other	10	0.10	ECAPA	0.00204	0.00178	0.00269	0.00074	0.00135	0.00254
Other	10	0.10	WavLM	0.00279	0.00252	0.00334	0.00840	0.01200	0.00254
Other	10	0.10	x-vector	0.00279	0.00252	0.00334	0.02271	0.03067	0.00254
Other	20	0.01	ECAPA	0.01782	0.01686	0.01780	0.01781	0.01781	0.00115
Other	20	0.01	WavLM	0.02475	0.02493	0.02375	0.01888	0.01121	0.00115
Other	20	0.01	x-vector	0.02475	0.02493	0.02375	0.02734	0.02182	0.00115
Other	20	0.10	ECAPA	0.00543	0.00543	0.00549	0.00394	0.00505	0.00254
Other	20	0.10	WavLM	0.00773	0.00773	0.00779	0.01621	0.02011	0.00254
Other	20	0.10	x-vector	0.00773	0.00773	0.00779	0.04748	0.12949	0.00254

A.1 COMPLETE ERROR-PLACEMENT INTERVENTION

We extended the intervention from the two featured contexts to all 24 corpus/noise/curator/prior conditions. With 20 initial corruption seeds and 20 permutations per seed, the grid contains 480 original placements and 9,600 permuted decisions. It reproduces all 840 outcomes from the two-context analysis exactly. Table 5 includes every condition.

Permutation reduces all 24 mean regrets, and 22 pointwise paired intervals exclude zero. Twenty comparisons have two-sided paired-t p-values below 0.05 after Holm adjustment across the 24-cell family. These approximate tests condition on the fixed speech frames and specified corruption mechanism. They neither quantify uncertainty over new speakers nor make this post-result extension confirmatory.

To test sensitivity to ties, we enumerate every globally proxy-optimal model–threshold policy in the frozen candidate and retained-score grid, using tolerance 10^{-12}. Within each initial seed, we subtract the mean of the 20 permutations' maximum tied reference regrets from the original placement's minimum tied regret. This conservative difference has a positive mean in all 24 conditions and positive pointwise intervals in 22. Only four of the 480 original placements admit different reference regrets across tied optima. The full-grid mean direction therefore survives every choice among proxy-optimal representatives in this policy set and tolerance.

Table 5: Complete matched-cohort permutation control. Entries are retained-frame normalized reference regrets after oracle cohort correction; original and permuted labels preserve class-specific error counts. Δ subtracts mean-permuted regret from original regret. Pointwise paired 95% t intervals use 20 initial seeds, and Holm p-values adjust all 24 two-sided contrasts. Inference remains conditional on these speaker frames.

Frame	Curator	Flip %	Prior	Original	Permuted	Δ [95% CI]	Holm p
Clean	ECAPA	10	0.01	0.00473	0.00002	0.00471 [0.00082, 0.00859]	0.0603
Clean	ECAPA	10	0.10	0.00294	0.00002	0.00292 [0.00152, 0.00432]	0.0046
Clean	ECAPA	20	0.01	0.00763	0.00009	0.00754 [0.00301, 0.01207]	0.0241
Clean	ECAPA	20	0.10	0.00713	0.00004	0.00709 [0.00527, 0.00891]	< 0.0001
Clean	x-vector	10	0.01	0.00683	0.00026	0.00657 [0.00194, 0.01120]	0.0465
Clean	x-vector	10	0.10	0.00603	0.00012	0.00591 [0.00351, 0.00831]	0.0010
Clean	x-vector	20	0.01	0.02804	0.00103	0.02700 [0.01458, 0.03942]	0.0033
Clean	x-vector	20	0.10	0.01612	0.00024	0.01588 [0.01023, 0.02154]	0.0002
Clean	WavLM	10	0.01	0.00840	0.00034	0.00807 [0.00357, 0.01256]	0.0175
Clean	WavLM	10	0.10	0.01269	0.00006	0.01263 [0.00714, 0.01812]	0.0021
Clean	WavLM	20	0.01	0.02639	0.00064	0.02576 [0.01430, 0.03721]	0.0025
Clean	WavLM	20	0.10	0.07471	0.00018	0.07453 [0.06239, 0.08666]	< 0.0001
Other	ECAPA	10	0.01	0.00357	0.00007	0.00351 [0.00104, 0.00597]	0.0465
Other	ECAPA	10	0.10	0.00066	0.00001	0.00066 [-0.00032, 0.00164]	0.3305
Other	ECAPA	20	0.01	0.00643	0.00024	0.00618 [0.00219, 0.01017]	0.0335
Other	ECAPA	20	0.10	0.00418	0.00008	0.00411 [0.00146, 0.00675]	0.0335
Other	x-vector	10	0.01	0.00243	0.00108	0.00135 [-0.00061, 0.00331]	0.3305
Other	x-vector	10	0.10	0.00913	0.00058	0.00855 [0.00563, 0.01147]	0.0001
Other	x-vector	20	0.01	0.00912	0.00231	0.00682 [0.00300, 0.01063]	0.0175
Other	x-vector	20	0.10	0.07041	0.00125	0.06915 [0.05368, 0.08463]	< 0.0001
Other	WavLM	10	0.01	0.00484	0.00105	0.00379 [0.00090, 0.00668]	0.0517
Other	WavLM	10	0.10	0.00523	0.00030	0.00493 [0.00215, 0.00771]	0.0175
Other	WavLM	20	0.01	0.01069	0.00301	0.00768 [0.00308, 0.01227]	0.0241
Other	WavLM	20	0.10	0.01014	0.00080	0.00935 [0.00575, 0.01295]	0.0006

Original-frame consequences of the recorded policies. The post-review extension evaluates each recorded model and threshold on its complete original reference frame. It reuses the saved selections and permutation draws; every original-placement cost matches the earlier oracle-cohort correction output. Table 6 reports the full additional outcome with the same pairing by initial seed. Mean regret falls in every clean-speech condition and all but one other-speech condition. A retained-frame improvement therefore need not carry over to the original population.

Table 6: Original-frame regret of policies from the matched-error intervention. Each model and threshold is selected on the retained cohort using its own oracle confusion correction, then evaluated on the complete original frame at the stated prior. The active curator is excluded from selection and the reference oracle. Means use 20 initial corruption seeds. Δ subtracts mean regret over 20 within-class permutations per seed from original-placement regret; intervals are pointwise paired 95% t intervals. Holm p adjusts the 24 added two-sided contrasts. Joint adjustment across all 48 retained- and original-frame contrasts preserves the same 13 original-frame rejections. This post-review exploratory outcome applies to the recorded policy representatives; tie invariance on the original frame is not established.

Frame	Curator	Flip %	p	Original	Permuted	Δ [95% CI]	Holm p
Clean	ECAPA	10	0.01	0.02458	0.01214	0.01244 [-0.00228, 0.02716]	0.6507
Clean	ECAPA	10	0.10	0.00854	0.00272	0.00582 [0.00291, 0.00874]	0.0079
Clean	ECAPA	20	0.01	0.02704	0.01118	0.01585 [0.00111, 0.03060]	0.3285
Clean	ECAPA	20	0.10	0.01750	0.00263	0.01487 [0.01126, 0.01847]	< 0.0001
Clean	x-vector	10	0.01	0.03907	0.02160	0.01747 [0.00223, 0.03270]	0.2681
Clean	x-vector	10	0.10	0.01123	0.00253	0.00870 [0.00370, 0.01370]	0.0225
Clean	x-vector	20	0.01	0.05718	0.02104	0.03614 [0.01811, 0.05417]	0.0079
Clean	x-vector	20	0.10	0.02355	0.00266	0.02090 [0.01244, 0.02936]	0.0010
Clean	WavLM	10	0.01	0.04044	0.02043	0.02001 [0.00032, 0.03970]	0.3735
Clean	WavLM	10	0.10	0.03160	0.00223	0.02937 [0.01773, 0.04101]	0.0008
Clean	WavLM	20	0.01	0.06552	0.01800	0.04752 [0.02477, 0.07028]	0.0056
Clean	WavLM	20	0.10	0.13822	0.00306	0.13516 [0.11391, 0.15640]	< 0.0001
Other	ECAPA	10	0.01	0.01039	0.00577	0.00461 [-0.00122, 0.01045]	0.6507
Other	ECAPA	10	0.10	0.00135	0.00088	0.00046 [-0.00100, 0.00193]	1.0000
Other	ECAPA	20	0.01	0.01781	0.01157	0.00625 [-0.00141, 0.01391]	0.6507
Other	ECAPA	20	0.10	0.00505	0.00113	0.00392 [0.00080, 0.00705]	0.1830
Other	x-vector	10	0.01	0.00568	0.00757	-0.00189 [-0.00698, 0.00320]	1.0000
Other	x-vector	10	0.10	0.03067	0.00664	0.02403 [0.01752, 0.03055]	< 0.0001
Other	x-vector	20	0.01	0.02182	0.01204	0.00978 [0.00407, 0.01548]	0.0236
Other	x-vector	20	0.10	0.12949	0.00712	0.12237 [0.09531, 0.14943]	< 0.0001
Other	WavLM	10	0.01	0.00478	0.00371	0.00107 [-0.00178, 0.00392]	1.0000
Other	WavLM	10	0.10	0.01200	0.00362	0.00838 [0.00387, 0.01289]	0.0138
Other	WavLM	20	0.01	0.01121	0.00812	0.00309 [-0.00113, 0.00730]	0.6507
Other	WavLM	20	0.10	0.02011	0.00467	0.01544 [0.01106, 0.01982]	< 0.0001

B Population effects and decision conventions

Separating population and measurement changes. Write $g = 0$ for the unfiltered frame and $\Delta_{ij}^{L}(g)$ for a model-risk difference on cohort g, keeping the cost and optimization convention common across comparisons. The observed change decomposes as

$$\Delta_{ij}^{Z}(g) - \Delta_{ij}^{Z}(0) = \left[\Delta_{ij}^{Y}(g) - \Delta_{ij}^{Y}(0)\right] + \left[B_{ij}(g) - B_{ij}(0)\right]. \tag{11}$$

Here $B_{ij}(g) = \Delta_{ij}^{Z}(g) - \Delta_{ij}^{Y}(g)$ follows the chosen risk convention. The brackets separate reference-risk changes across populations from changes in measurement distortion. Retained-frame results alone still do not identify original-frame performance.

Finite threshold policies and ties. Thresholding scores s_m at t produces at most $N + 1$ distinct decision vectors, including always-accept and always-reject. Equal-score trials change decisions together, regardless of their reference labels. In the independent-pair study, we encode each transition with floating-point `nextafter` immediately above the corresponding score and explicitly include both infinite endpoints. The theoretical results assume exact minimization. Computation treats threshold policies whose costs lie within absolute 10^{-12} of the minimum as tied, choosing the earliest threshold in increasing order. The model with lowest computed cost is selected, with exact ties resolved by recorded model order; strict-winner and tied-optimum summaries apply the same 10^{-12} tolerance. Reporting both strict wrong-winner events and reference-cost ranges across proxy optima distinguishes a reversal from an ambiguous proxy choice. Appendix E specifies the model-tie and threshold rules for the identity and subset studies.

We compute EER by linearly interpolating the adjacent ROC vertices where $\mathrm{FNR} - \mathrm{FPR}$ crosses zero, equivalent to randomizing between neighboring threshold decisions. The curator requires a deterministic development policy, so its threshold minimizes $|\,\mathrm{FNR} - \mathrm{FPR}\,|$ over vertices. Reported EER and the curator's chosen threshold retain these distinct definitions.

C Model, data, and computational provenance

We evaluate five released checkpoints: SpeechBrain `spkrec-ecapa-voxceleb` and `spkrec-xvect-voxceleb`, Microsoft `wavlm-base-plus-sv`, and WeSpeaker `wespeaker-voxceleb-resnet34-LM` and `wespeaker-voxceleb-resnet221-LM`. Each comparison uses cosine similarity between normalized embeddings. Different architectures need not make independent errors. Architecture and toolkit papers establish the model-family lineage (Ravanelli et al., 2021; Desplanques et al., 2020; Snyder et al., 2018; Chen et al., 2022; Wang et al., 2023); the experiment records exact releases, model-card or pretrained-registry links, snapshot revisions, and file hashes.[1]

Table 7: Frozen Hugging Face snapshots, shown as 12-character revision prefixes. Manifests retain full revisions and weight hashes.

Checkpoint	Revision
ECAPA	`0f99f2d0ebe8`
x-vector	`56895a2df401`
WavLM	`feb593a6c23c`
ResNet34	`f0c48c298fd8`
ResNet221	`26732f43821f`

We use mono 16 kHz FLAC recordings lasting at least two seconds and retain at most the first six seconds. A speaker qualifies with at least four usable utterances in each of two distinct chapters. We select the two chapters with the most eligible utterances, resolve chapter ties numerically, and take each chapter's first four eligible utterances in numeric order. One chapter provides enrollment recordings and the other provides probes. The trial list contains every enrollment–probe combination; no recording serves both roles. This rule yields 31 development and 29 test speakers on clean

[1] Checkpoint releases: ECAPA, x-vector, WavLM, ResNet34-LM, and ResNet221-LM. The official pretrained-model registry documents WeSpeaker training provenance.

speech, and 28 development and 29 test speakers on other speech. The four speaker sets are mutually disjoint. Across 936 distinct recording inputs, the development frames contain 15,376 and 12,544 trials; each test frame contains 13,456 trials, of which 464 are same-speaker pairs.

Inference follows the released SpeechBrain feature and encoder pipelines and the released WavLM feature extractor and x-vector output. WeSpeaker's official ONNX checkpoints receive 80-bin filterbanks with 25 ms frames, 10 ms shifts, Hamming windows, zero dither, PCM scaling by 32768 and utterance mean normalization. No model is trained. The recorded software versions are PyTorch/TorchAudio 2.8.0, SpeechBrain 1.0.3, Transformers 4.57.1 and ONNX Runtime 1.30.0. Independent checks reconstruct the score matrices from cached embeddings and verify complete IDs, waveform hashes, finite scores, enrollment/probe separation, and reference labels. LibriSpeech metadata supply the development and test references (Panayotov et al., 2015). Libri-Light excludes LibriSpeech development/test speakers (Kahn et al., 2019). This establishes separation for that pretraining component only; overlap with every training source has not been ruled out.

The checkpoint panel, frozen score matrices, and earlier analysis records are retained in the reproduction package. The correction, permutation, structured-corruption, and subset-selection extensions are exploratory; none was externally preregistered.

We flip pair labels with probabilities 0.10 and 0.20 using seeds 260926–260945, pairing each seed across gates and settings. Equal test-frame sizes allow the same flip-index draws, which also couple the two frames' conditional corruption results. A separate control retains clean labels. Each curator gates at its own threshold nearest the development equal-error point, retaining $Z = c$. Diagnostic outputs include the curator; the primary choice set excludes it and contains four models. We set $C_M = C_F = 1$ and priors 0.01 and 0.10.

The reference-optimal test model and threshold provide a finite-frame oracle comparator. Original-frame regret subtracts the minimum eligible normalized reference cost from the selected policy's cost on that same complete trial frame. It does not measure a separate deployment holdout; future speakers or traffic require another evaluation population. We omit nominal trial-level confidence intervals over the reused recordings. The reported paired-corruption intervals and subset-sampling summaries have the conditional interpretations stated in the tables.

D COMPLETE CONTROLLED-STUDY SUMMARIES

Cleaner labels can support worse decisions. With 10% imposed noise on clean speech, WavLM agreement retention lowers mean label error from 10.00% to 0.983%, yet raises mean original-frame regret at prior 0.10 from 0.00509 to 0.02057. The paired change is 0.01547, with pointwise 95% interval $[0.00898, 0.02197]$. On other speech, x-vector retention leaves 1.326% label error but raises regret from 0.01073 to 0.02271, a difference of 0.01198 with interval $[0.00333, 0.02062]$. The complete tables below retain every other curator/frame condition, including gains and intervals spanning zero. Both illustrations were selected after the full grid was inspected, rather than specified as confirmatory endpoints.

The direction depends on the operating tradeoff: at prior 0.01, curation lowers mean original-frame regret in every curator/noise/frame condition. Clean WavLM at 10% noise, for example, drops from 0.81903 to 0.01907. Curation also improves five of the six prior-0.10 conditions at 20% noise. These gains rule out a blanket interpretation that filtering harms evaluation.

Ranking, calibration, and population effects differ. At prior 0.10, the clean WavLM example has retained-frame regret 0.00780, comprising 0.00119 from model choice and 0.00661 from threshold choice; the latter accounts for 84.8% of the mean total. Threshold choice accounts for 71.9% in the other-speech x-vector example. Evaluating on the original frame additionally measures the effect of selecting a policy on a different score population. When proxy and reference labels coincide in the clean-label controls, retained regret is zero but original-frame regret can remain positive after curation. Label errors therefore cannot explain the entire operational outcome.

Count a strict wrong winner only when the unique proxy-optimal model lies outside the reference-optimal set; a tied proxy optimum is a non-event. This count changes with the candidate panel: curated clean speech produces no such event among the original three checkpoints but does among the expanded panel. Additional checks vary the choice among tied optima and remove all trials involving each test speaker in turn; these checks are dependent. Reporting decision cost alongside the counts shows whether an ordering change has a material consequence.

D.1 CLEAN SPEAKER PANEL

Table 8: Clean panel: complete original-frame regret comparisons before and after agreement filtering. Means use 20 imposed-corruption seeds; pointwise 95% paired t intervals summarize mean corruption effects conditional on the fixed speakers and scores. Conditions are dependent, and intervals have no multiplicity adjustment.

Flip %	Curator	p	Before	After	Δ	95% CI
10	ECAPA	0.01	0.8393	0.0246	-0.8147	[-0.8649,-0.7645]
10	ECAPA	0.10	0.0051	0.0074	+0.0023	[-0.0009,+0.0055]
10	WavLM	0.01	0.8190	0.0191	-0.8000	[-0.8719,-0.7280]
10	WavLM	0.10	0.0051	0.0206	+0.0155	[+0.0090,+0.0220]
10	x-vector	0.01	0.7995	0.0365	-0.7631	[-0.8339,-0.6923]
10	x-vector	0.10	0.0051	0.0086	+0.0036	[-0.0007,+0.0078]
20	ECAPA	0.01	0.9484	0.0247	-0.9238	[-0.9382,-0.9093]
20	ECAPA	0.10	0.0563	0.0168	-0.0395	[-0.0691,-0.0099]
20	WavLM	0.01	0.9490	0.0441	-0.9049	[-0.9223,-0.8875]
20	WavLM	0.10	0.0571	0.0604	+0.0033	[-0.0334,+0.0399]
20	x-vector	0.01	0.9488	0.0435	-0.9053	[-0.9162,-0.8943]
20	x-vector	0.10	0.0571	0.0141	-0.0430	[-0.0691,-0.0170]

Table 9: Clean panel: retained-label quality and retained-frame regret decomposition at prior 0.10. A wrong-winner event requires a unique proxy winner outside the reference-optimal set. Tied proxy optima count as no event, and frequencies use all 20 seeds. Model and threshold components sum to regret on the retained frame rather than the original frame.

Flip %	Curator	Keep %	Error %	Wrong %	Model	Threshold
10	ECAPA	88.76	0.171	20	0.0008	0.0016
10	x-vector	85.33	0.688	45	0.0027	0.0019
10	WavLM	83.64	0.983	10	0.0012	0.0066
20	ECAPA	79.09	0.381	55	0.0022	0.0046
20	x-vector	76.50	1.513	60	0.0041	0.0056
20	WavLM	75.24	2.132	30	0.0022	0.0300

D.2 Other speaker panel

Table 10: Other panel: complete original-frame regret comparisons before and after agreement filtering. Means use 20 imposed-corruption seeds; pointwise 95% paired t intervals summarize mean corruption effects conditional on the fixed speakers and scores. Conditions are dependent, and intervals have no multiplicity adjustment.

Flip %	Curator	p	Before	After	Δ	95% CI
10	ECAPA	0.01	0.7387	0.0104	-0.7283	[-0.7812,-0.6755]
10	ECAPA	0.10	0.0083	0.0007	-0.0076	[-0.0101,-0.0050]
10	WavLM	0.01	0.7360	0.0075	-0.7286	[-0.7895,-0.6676]
10	WavLM	0.10	0.0107	0.0084	-0.0023	[-0.0073,+0.0026]
10	x-vector	0.01	0.7555	0.0076	-0.7479	[-0.8022,-0.6937]
10	x-vector	0.10	0.0107	0.0227	+0.0120	[+0.0033,+0.0206]
20	ECAPA	0.01	0.8773	0.0178	-0.8595	[-0.8696,-0.8493]
20	ECAPA	0.10	0.0687	0.0039	-0.0647	[-0.0781,-0.0514]
20	WavLM	0.01	0.8751	0.0189	-0.8563	[-0.8670,-0.8455]
20	WavLM	0.10	0.0680	0.0162	-0.0518	[-0.0680,-0.0355]
20	x-vector	0.01	0.8793	0.0273	-0.8520	[-0.8615,-0.8425]
20	x-vector	0.10	0.0680	0.0475	-0.0205	[-0.0406,-0.0004]

Table 11: Other panel: retained-label quality and retained-frame regret decomposition at prior 0.10. A wrong-winner event requires a unique proxy winner outside the reference-optimal set. Tied proxy optima count as no event, and frequencies use all 20 seeds. Model and threshold components sum to regret on the retained frame rather than the original frame.

Flip %	Curator	Keep %	Error %	Wrong %	Model	Threshold
10	ECAPA	89.02	0.127	0	0.0000	0.0002
10	x-vector	81.52	1.326	60	0.0018	0.0046
10	WavLM	85.07	0.750	20	0.0008	0.0018
20	ECAPA	79.31	0.304	0	0.0000	0.0032
20	x-vector	73.62	2.879	85	0.0053	0.0157
20	WavLM	76.34	1.645	50	0.0028	0.0044

D.3 Sensitivity and complete outputs

The reproduction archive contains every model's proxy/reference EER and minDCF, signed optimized bias, selected thresholds, fixed-threshold distortion, calibration regret, and all six eligible evaluator-pair comparisons per five-model condition. Its machine-readable CSVs preserve the original three-model subset, every corruption seed, model/threshold tie ranges, and leave-one-speaker-out checks. Comparisons that share models or trials are dependent and are not counted as independent experiments.

At 10% imposed noise and prior 0.10, mean original-frame regret for clean WavLM curation is 0.02057 under the stated tie rule. Averaging each seed's minimum and maximum regret across all proxy-optimal policies gives 0.01749 and 0.02071. For other x-vector curation, the corresponding values are 0.02271, 0.02226 and 0.02282. These are ranges over admissible tie choices, not sampling confidence intervals. Positive regret in these examples therefore persists beyond the arbitrary choice of a tied model winner.

Panel composition changes wrong-winner counts. On the clean frame, the original three-model subset has no strict curated wrong-winner event across noisy settings; the five-model panel has 79 among 240 dependent curator/noise/prior/seed cells after agreement retention. The other frame has 6 of 240 and 72 of 240, respectively. These summarize panel sensitivity rather than independent hypothesis tests. Each candidate set also defines its own reference oracle, changing the scale of absolute regret.

Setting $Z = Y$ before agreement filtering removes retained-frame label-induced regret. At prior 0.10, original-frame regrets nevertheless remain 0.00331, 0.00139 and 0.00208 for clean ECAPA, x-vector and WavLM gates, and 0.00054, 0.00246 and 0.00054 on the other frame. Population selection alone can therefore change the chosen policy's cost. Original-frame outcomes under noisy labels combine this effect with label error.

E Structured identities and provisional subset protocols

Identity partitions. Each test frame has 29 speakers with eight recordings each. For seeds 290926–290945, we construct one provisional identity assignment per recording and use equality of those assignments to label every enrollment–probe pair. Low severity merges two disjoint speaker pairs and splits two other speakers; high severity uses five of each, with the low changes nested in the high construction. Each split randomly assigns four of a speaker's eight recordings a new identity. The assignment is reused throughout the frame, including whenever a recording appears in multiple trials.

The random mechanism draws speaker pairs and split speakers from one fixed-seed permutation. The similarity-based mechanism ranks unordered speaker pairs by their mean full-target ECAPA score over the 32 cross-speaker trials in both enrollment directions, breaks ties by numeric speaker IDs, and greedily takes the first five disjoint pairs. These merges are fixed across seeds; split speakers and recording assignments vary. ECAPA defines this imposed dependence even when another model is the curator. Reference groups and source scores construct the simulation; they are not supplied as extra information to unadjusted policy selection.

For each partition, the matched pair control uniformly relocates its exact numbers of false-negative and false-positive labels within their respective reference classes. This control preserves both class-specific error counts and proxy prevalence but generally violates identity consistency. It is a mechanism counterfactual, not a proposed metadata repair. Curation is reapplied separately, so the two retained populations need not match.

Identity-grid decisions and uncertainty. The gates use the frozen trusted-development thresholds. Full and retained policies either minimize unadjusted proxy DCF or use their own oracle confusion counts for correction. All comparisons exclude the active curator. Retained selection enumerates the original score grid, preserving distinct original-frame decisions inside gaps left by filtering; every global optimum within 10^{-12} contributes to the recorded tie envelope. The chosen representative uses the earliest eligible model and then increasing threshold order, without looking at reference regret. Development-only choices use their development grid. Tables 12–13 retain every identity condition; machine-readable outputs additionally retain matched controls, 20-seed paired differences, pointwise intervals, and tie envelopes. These intervals describe the declared randomization on fixed frames. Fixed near-merge pairs can yield identical selected policies across seeds; zero empirical seed variation is not population certainty.

Subset construction. The budget study uses independent pair flips with probabilities 0.10 and 0.20 and seeds 260926–260945. The selector's arguments are Z, at most one curator-score vector, the row budget, the method, and a random generator. It cannot inspect speaker IDs, reference labels, candidate responses, or recording endpoints. Each budget/method/curator/frame/noise/seed combi-

nation uses one separately seeded subset draw, reused across priors. There are 2160 selections and 4320 fitted decisions, with no absent-class fits or budget shortfalls.

Uniform sampling draws without replacement from the complete frame with inclusion probability b/N. Agreement sampling first minimizes the curator's absolute difference between provisional false-accept and false-reject rates over its score vertices, takes the earliest increasing threshold, and samples uniformly where that decision equals Z. Its inclusion probability is b/N_{agree} inside the pool and zero outside. The coverage design forms empirical score octiles using linear quantiles, assigns boundary ties to the upper bin, and crosses the bins with Z. Each nonempty stratum receives one draw; remaining slots repeatedly go to the least-filled noncensus stratum, with index-order tie breaking. Sampling is uniform without replacement inside each stratum, giving $\pi_i = n_h/N_h > 0$.

Candidate scores are revealed only for sampled rows and the four eligible models. Weighted positive, negative, miss, and false-accept totals use $1/\pi_i$; their class-conditional ratios define proxy DCF. For coverage sampling, each stratum has one proxy class and known size, so the weighted class denominators equal the full proxy totals exactly. Horvitz–Thompson unbiasedness then applies to the cost of any fixed policy, but does not make its optimized minimum or selected policy unbiased. Uniform sampling has random class denominators; agreement sampling targets its retained pool. The labels remain binary and the sampling weights positive, so fitted curves are valid monotone empirical rates whenever both classes appear.

Subset thresholds and evaluation. Policy fitting uses only distinct selected scores, keeping equal scores together and including both constant policies. Finite transitions use `nextafter`; the last transition uses $+\infty$. The earliest threshold within cost tolerance 10^{-12} and then the earliest eligible model resolve ties. This fixes behavior on unseen scores without consulting them; it does not establish invariance over all possible thresholds in unseen score gaps. The separate validation harness evaluates that recorded policy on the full provisional and reference frames and uses each frame's own eligible optimum. All published budget means are descriptive equal-weight averages over the stated grid. The complete condition summaries and paired contrasts remain in the reproduction package.

Processing ledger. For b selected pairs and $U(T)$ unique endpoints out of $U = 232$ recordings, uniform sampling uses $4b$ candidate scalar scores and $4U(T)$ evaluated-encoder recording passes. A score-aware method adds $N = 13456$ curator scores and U curator recording passes. Selection overhead, pair scoring, and architecture-specific inference time are not interchangeable. The cached-score experiment measures no runtime speedup; it reports information accesses and recording coverage, with the offline full-score/reference evaluation treated as research overhead. Different methods at the same row budget need not have equal computational cost.

Why scores and provisional labels cannot certify truth. Consider four recordings and two possible reference partitions, $\{1,2\}|\{3,4\}$ and $\{1,3\}|\{2,4\}$. Two candidate score vectors indicate equality under the respective partitions on all six unordered pairs. Hold the observed provisional partition fixed at $\{1,4\}|\{2,3\}$. Either reference world has both classes, but a different candidate has zero reference cost. The selector receives identical scores and provisional labels in the two worlds. Without additional identifying assumptions it cannot determine which reference conclusion is correct, even when every partition is globally consistent. This elementary ambiguity limits what provisional coverage or stability can establish.

Table 12: Clean identity-consistent grid: mean original-frame normalized reference regret over 20 seeds. Each curator is excluded from its candidate set. Corrected uses each evaluated population's own oracle confusion counts. Low/high is two/five merges plus two/five splits. All cells share the fixed acoustic frame and are dependent.

				Unadjusted		Corrected	
Merges	Level	Curator	p	Full	Retained	Full	Retained
ECAPA-near	High	ECAPA	0.01	0.07037	0.01342	0.01088	0.01342
ECAPA-near	High	ECAPA	0.10	0.00608	0.02063	0.15979	0.02063
ECAPA-near	High	WavLM	0.01	0.06291	0.15980	0.08755	0.16449
ECAPA-near	High	WavLM	0.10	0.05614	0.16048	0.07120	0.21382
ECAPA-near	High	x-vector	0.01	0.04338	0.15827	0.08755	0.15827
ECAPA-near	High	x-vector	0.10	0.05614	0.06283	0.07120	0.06635
ECAPA-near	Low	ECAPA	0.01	0.00543	0.01397	0.00404	0.01397
ECAPA-near	Low	ECAPA	0.10	0.00608	0.02063	0.01204	0.02063
ECAPA-near	Low	WavLM	0.01	0.03557	0.11957	0.03557	0.13341
ECAPA-near	Low	WavLM	0.10	0.02293	0.07666	0.02366	0.07666
ECAPA-near	Low	x-vector	0.01	0.03557	0.13215	0.03557	0.15027
ECAPA-near	Low	x-vector	0.10	0.02293	0.02366	0.02366	0.02827
Random	High	ECAPA	0.01	0.02268	0.01397	0.00713	0.01397
Random	High	ECAPA	0.10	0.00055	0.00405	0.00134	0.00405
Random	High	WavLM	0.01	0.03856	0.00453	0.00713	0.00453
Random	High	WavLM	0.10	0.00109	0.00467	0.00128	0.00467
Random	High	x-vector	0.01	0.03856	0.03645	0.00713	0.03645
Random	High	x-vector	0.10	0.00109	0.00406	0.00128	0.00406
Random	Low	ECAPA	0.01	0.00596	0.01424	0.00522	0.01424
Random	Low	ECAPA	0.10	0.00007	0.00312	0.00007	0.00312
Random	Low	WavLM	0.01	0.00784	0.00115	0.00522	0.00115
Random	Low	WavLM	0.10	0.00007	0.00208	0.00007	0.00208
Random	Low	x-vector	0.01	0.00784	0.03091	0.00522	0.03091
Random	Low	x-vector	0.10	0.00007	0.00213	0.00007	0.00213

Table 13: Other identity-consistent grid: mean original-frame normalized reference regret over 20 seeds. Each curator is excluded from its candidate set. Corrected uses each evaluated population's own oracle confusion counts. Low/high is two/five merges plus two/five splits. All cells share the fixed acoustic frame and are dependent.

				Unadjusted		Corrected	
Merges	Level	Curator	p	Full	Retained	Full	Retained
ECAPA-near	High	ECAPA	0.01	0.01919	0.04094	0.02408	0.04094
ECAPA-near	High	ECAPA	0.10	0.00179	0.00062	0.00536	0.00062
ECAPA-near	High	WavLM	0.01	0.02881	0.00843	0.02408	0.00765
ECAPA-near	High	WavLM	0.10	0.00388	0.00698	0.10170	0.00826
ECAPA-near	High	x-vector	0.01	0.02881	0.03564	0.02408	0.03462
ECAPA-near	High	x-vector	0.10	0.00388	0.18180	0.10170	0.19220
ECAPA-near	Low	ECAPA	0.01	0.00475	0.00115	0.00146	0.00115
ECAPA-near	Low	ECAPA	0.10	0.00000	0.00055	0.00000	0.00055
ECAPA-near	Low	WavLM	0.01	0.00709	0.00222	0.00241	0.00220
ECAPA-near	Low	WavLM	0.10	0.01385	0.01479	0.01385	0.01479
ECAPA-near	Low	x-vector	0.01	0.00709	0.00283	0.00241	0.00283
ECAPA-near	Low	x-vector	0.10	0.01385	0.06135	0.01385	0.06135
Random	High	ECAPA	0.01	0.03300	0.00252	0.00680	0.00252
Random	High	ECAPA	0.10	0.00056	0.00109	0.00169	0.00109
Random	High	WavLM	0.01	0.04494	0.00438	0.00754	0.00364
Random	High	WavLM	0.10	0.00056	0.00330	0.00169	0.00330
Random	High	x-vector	0.01	0.04494	0.00434	0.00754	0.00434
Random	High	x-vector	0.10	0.00056	0.01033	0.00169	0.01033
Random	Low	ECAPA	0.01	0.00643	0.00252	0.00319	0.00252
Random	Low	ECAPA	0.10	0.00018	0.00055	0.00018	0.00055
Random	Low	WavLM	0.01	0.00643	0.00179	0.00319	0.00207
Random	Low	WavLM	0.10	0.00018	0.00116	0.00018	0.00116
Random	Low	x-vector	0.01	0.00643	0.00434	0.00319	0.00434
Random	Low	x-vector	0.10	0.00018	0.00479	0.00018	0.00489