CERTIFIED DECISION TRANSFER UNDER UNLABELED SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

In content moderation and fraud detection, model upgrades should identify more positive cases without increasing review workload. Yet higher offline AUC does not guarantee this improvement under distribution shift, delayed target labels hinder evaluation before deployment, and changes in score scale complicate threshold reuse. We study *certified decision transfer*: using historical labels and current unlabeled traffic to decide whether a candidate should replace an existing policy. Rate Matching (RM) lets the reference determine how many cases to select and the candidate determine which cases fill those slots. We derive finite-batch and population F_β regret decompositions that separate selection-count error from ranking error, and show why global AUC superiority is insufficient at a fixed review count. Certified RM (C-RM) uses paired policy disagreements on historical positives and bounds target selection-rate changes using unlabeled traffic. For pre-specified policies evaluated on independent certification samples, stable positive-case detection rates enable finite-sample control of erroneous approvals for each comparison, simultaneously for all $\beta > 0$; insufficient evidence retains the reference. In a GoEmotions audit, 3,578 of 26,250 class-level upgrade comparisons are harmful; fully paired C-RM approves 834 with no observed population or finite-batch F1 degradation. On COCO, disagreement factorization certifies 4.0–4.6 times as many upgrades as the direct-cell test using the same data and nominal error budget. Our framework supports evidence-based model replacement before target labels become available. Code is available at `https://anonymous.4open.science/r/CDT-86F8/`.

1 INTRODUCTION

In content moderation and fraud detection, an improved model should help reviewers find more violations or fraudulent transactions within the same workload. A higher AUC on historical data does not establish that it will do so. AUC measures ranking across all score thresholds, whereas a review queue depends on the cases selected at one operating point. Current traffic may also differ from the data used to select the model. When investigation or review outcomes arrive days later, the evidence needed to evaluate an upgrade on that traffic is unavailable at the time of the deployment decision.

Even preserving the workload can be difficult when models use different score scales: applying the old threshold to a new model may select substantially more or fewer cases. We study replacement relative to an existing reference policy. If that policy selects 120 of 10,000 transactions, the candidate can rank the same transactions and supply its top 120 for review. This rule, which we call Rate Matching (RM), transfers the reference's selection count while letting the candidate determine the selected cases. It uses classical Classify-and-Count (Tasche, 2016; Schumacher et al., 2025) to set the count and requires no calibration between the two models' score scales. Approving the replacement then requires evidence that the candidate selects more positive cases at the same workload, despite the absence of target labels.

Certified Rate Matching (C-RM) tests this replacement using independent historical positives and unlabeled target traffic. A matching window sets the candidate cutoff; certification then compares the resulting fixed policies. Under stability of their positive-case detection rates, C-RM controls

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.

erroneous approval of a population F_β regression, simultaneously for all $\beta > 0$. An approved candidate serves subsequent traffic; otherwise, the reference continues to make the decisions.

The test focuses on changed decisions: positives selected by both policies cancel in the capture-rate difference. Historical positives measure gains and losses, while unlabeled traffic measures selection-rate changes. Separating how often policies disagree from which policy benefits yields a more powerful test at the same nominal error level.

Experiments show how C-RM avoids harmful replacements and admits supported improvements. On GoEmotions, ModernBERT has higher source AUC than RoBERTa in all three training runs but lower target F1 in each; C-RM retains RoBERTa. Qwen improves on TF–IDF on average yet loses in 3,578 of 26,250 class-level comparisons. Fully paired C-RM approves 834 replacements with no observed population or finite-batch F1 degradation (Table 3). Local certification thus addresses regressions hidden by a better average score.

On COCO, factorization approves 4.0–4.6 times as many replacements with the same data and nominal error budget (Table 4). Natural-shift and evidence-budget audits further measure when available data support deployment.

Our contributions are:

- **Decision-level analysis of model replacement.** We formulate transfer at a reference-derived selection count and derive exact finite-batch and population F_β regret decompositions. The analysis identifies ranking quality at the selected depth as the relevant comparison, characterizes that depth through marginal precision, and constructs AUC reversals in which a globally better ranker makes worse local decisions.
- **A paired certificate under unlabeled shift.** We derive a sufficient condition for population F_β improvement that combines source-positive policy disagreements with unlabeled target selection rates. For fixed policies, independent certification samples, and stable positive-class detection rates, the resulting test provides finite-sample error control simultaneously for all $\beta > 0$. A complementary identifiability result shows that source labels and unlabeled target data alone cannot determine the ordering under unrestricted conditional shift.
- **Disagreement factorization for higher certification power.** We separate disagreement frequency from its direction and construct exact-binomial bounds on these quantities. On COCO, this yields 4.0–4.6 times as many certified upgrades as direct-cell estimation with the same data and nominal risk. Extensions incorporate specified drift budgets and family-wise error control for multiple comparisons.

2 Related Work

F-measure, quantification, and label shift. F-measure optimization learns decision rules for labeled distributions (Koyejo et al., 2014; Narasimhan et al., 2015). Our analysis instead compares two models at a reference-derived decision depth and separates count error from ranking error. Label-shift methods estimate the target prior and adjust decisions (Saerens et al., 2002; Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2020; Dussap et al., 2023). RM uses classical CC (Tasche, 2016; Schumacher et al., 2025) to transfer action volume, not to recover target prevalence. Its ranking step is invariant to strictly increasing candidate-score transformations. This distinction matters when the action budget is set by an existing policy rather than unconstrained target F-measure maximization.

Model evaluation before target outcomes arrive. Related work studies target-label-budgeted model selection (Okanovic et al., 2025), unlabeled prediction-set adaptation (Kasa et al., 2025), and testable learning under shift (Klivans et al., 2024). Confidence- or disagreement-based estimators can predict a model's scalar target accuracy, but unlabeled OOD accuracy is not identified without assumptions (Garg et al., 2022). C-RM tests a different quantity: the sign of a paired local F_β change at a transported action depth. Whereas target-label-budgeted selectors choose a global model using target outcomes, C-RM evaluates a pre-specified replacement before those outcomes arrive.

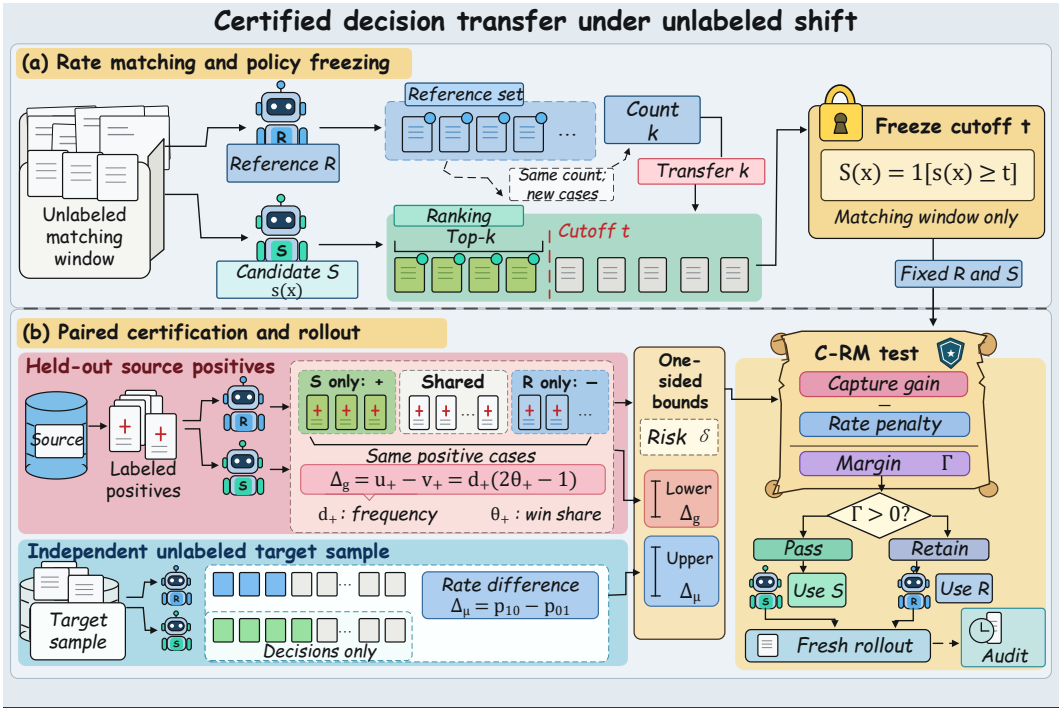

Figure 1: **From rate matching to certified replacement.** (a) The reference sets a count on an unlabeled matching window; the candidate fills those slots and freezes its cutoff. (b) Independent historical positives measure paired gains and losses, while unlabeled target cases measure selection-rate changes. Confidence bounds determine whether to deploy the frozen candidate or retain the reference. Count matching is exact on the matching window; certification concerns population improvement under positive-class stability. Delayed rollout labels audit the decision.

Calibration and statistical risk control. Risk-controlling prediction sets use labeled calibration data to bound prediction risk (Bates et al., 2021). Learn then Test formulates calibration as a multiple-testing problem (Angelopoulos et al., 2021), while Conformal Risk Control bounds the expected value of a monotone loss under exchangeable calibration (Angelopoulos et al., 2022). C-RM follows the separation of policy construction from statistical testing, but its target is relative population F_β improvement between fixed policies under source-to-target shift. Historical positive labels estimate capture differences; unlabeled target observations estimate selection rates, linked through positive-class stability. Risk-controlled post-processing (Joshi et al., 2026) calibrates on labeled target pairs to control pointwise violation risk. We therefore report its target-labeled diagnostic separately from methods that use no target labels, and distinguish absolute risk control from relative F_β improvement.

Conservative policy and model updates. High-confidence policy improvement and baseline bootstrapping likewise retain a baseline when evidence for an update is insufficient (Thomas et al., 2015; Laroche et al., 2019). Their guarantees concern policies in batch reinforcement learning with logged rewards or transition data. C-RM adapts the conservative-update principle to supervised decision policies with delayed target outcomes, incompatible score scales, a reference-derived action volume, and an F_β objective. Model-update work instead reduces negative flips or embedding incompatibility during training (Yan et al., 2021; Shen et al., 2020). Those methods are complementary: C-RM is a model-agnostic deployment-time gate and does not constrain how the receiver is trained.

3 DECISION-TRANSFER SETUP AND METHOD

We consider a multilabel decision problem with classes $c = 1, \ldots, C$ and describe the method for one class. Let P and Q denote the historical source and current target distributions, with positive-class priors $p = P(Y = 1)$ and $q = Q(Y = 1) > 0$. A labeled source validation set of size m supports model selection and threshold calibration. The deployed *reference* policy is $R(x) = \mathbf{1}[r(x) \geq \tau]$, where r is its score and τ its source-calibrated threshold; write $A = \{x : R(x) = 1\}$. A *candidate* model or ensemble provides scores $s(x)$, whose numerical scale may differ from that of r. Given unlabeled target traffic, our task is to construct a candidate decision policy at the reference workload and decide whether to deploy it using independent historical labels and current unlabeled observations. Figure 1 summarizes policy construction, independent certification, and deployment.

Rate matching. On a target matching window $D_t = \{x_i\}_{i=1}^n$, the reference determines

$$\widehat{k}_{RM} = \sum_{i=1}^{n} \mathbf{1}[r(x_i) \geq \tau]. \tag{1}$$

The candidate selects the top-$\widehat{k}_{RM}$ examples ranked by s. Their identities can change while the count stays fixed. Selections are invariant to strictly increasing transformations of s: the reference supplies the workload and the candidate supplies the ordering.

Fixing the proposed replacement. C-RM turns this batch-level match into a policy that can be evaluated before deployment. On the matching window, choose a cutoff t that selects the candidate's top-$\widehat{k}_{RM}$ cases and fix $S(x) = \mathbf{1}[s(x) \geq t]$. Score ties use a pre-specified secondary ordering, carried into the fixed policy; the notation suppresses this tie-breaker. Both R and S are fixed before certification. Their counts agree on the matching window, while their population selection rates $\mu_R = Q(R = 1)$ and $\mu_S = Q(S = 1)$ may differ. Independent target observations measure this difference for the certificate.

Paired evidence for the change. The comparison centers on cases selected by only one policy. On held-out source positives, define

$$u_+ = P(S = 1, R = 0 \mid Y = 1), \qquad v_+ = P(S = 0, R = 1 \mid Y = 1).$$

These are the positive cases gained and lost by the candidate. Certification assumes that each fixed policy's positive-class capture rate transfers from source to target: $P(H = 1 \mid Y = 1) = Q(H = 1 \mid Y = 1)$ for $H \in \{R, S\}$. Writing $g_H = Q(H = 1 \mid Y = 1)$, the target capture gain is then $\Delta_g = g_S - g_R = u_+ - v_+$. On independent unlabeled target traffic, define $p_{10} = Q(S = 1, R = 0)$ and $p_{01} = Q(S = 0, R = 1)$; their difference is the selection-rate change $\Delta_\mu = \mu_S - \mu_R = p_{10} - p_{01}$. Thus historical positives supply evidence about additional true cases, while current traffic supplies evidence about the change in workload.

Certification and deployment. For a fixed policy H, population $F_{\beta,H} = (1 + \beta^2)qg_H/(\beta^2 q + \mu_H)$. C-RM combines a lower bound on capture gain with an upper bound on selection-rate increase in the paired margin of Equation 5. The direct-cell construction bounds each disagreement direction separately. The factorized construction instead estimates disagreement frequency and the candidate-only share among disagreements, increasing power with the same samples and risk budget (Proposition 1).

A positive margin deploys S; otherwise deployment keeps R. Certification uses samples independent of model selection and matching, and deployment draws fresh traffic from the certification target distribution. The cutoff remains fixed throughout this deployment. Appendix A gives the algorithm and deployment timeline. Section 4 derives the paired certificate and its bounded-drift extension; the count-transport analysis additionally considers pure label shift, $Q(X \mid Y) = P(X \mid Y)$.

4 DECISION-REGRET AND STATISTICAL ANALYSIS

The analysis follows the deployment decision in Section 3. We first explain what changing the selected cases can improve at a fixed workload. We then derive the approval condition and show how

disagreement factorization uses the available evidence. Supporting results and proofs are collected in Appendix D.

4.1 What changes at a fixed workload

Let $N_+ > 0$ be the number of positives in a target batch and $TP_s(k)$ the number among the top-k cases under score s. At a common selection count,

$$F_{\beta,s}(k) - F_{\beta,r}(k) = \frac{(1+\beta^2)\{TP_s(k) - TP_r(k)\}}{\beta^2 N_+ + k}. \tag{2}$$

The denominator is shared: a replacement improves the decision precisely when it finds more positives within the same k slots. This is the local precision-at-depth condition that RM makes directly comparable across models with different score scales. AUC also counts ranking changes outside those slots, so a higher global AUC need not improve this decision. Appendix D.1 gives an explicit reversal construction and an analysis of how performance changes with selection depth.

The count and ranking play distinct roles. If $s^\star$ ranks all positives first, adding and subtracting its performance gives

$$1 - F_{\beta,s}(k) = \underbrace{1 - F_{\beta,s^\star}(k)}_{\text{selection-count error}} + \underbrace{F_{\beta,s^\star}(k) - F_{\beta,s}(k)}_{\text{ranking error at } k}. \tag{3}$$

Even a perfect ranker loses performance when too few or too many cases are selected; at a given count, the remaining loss comes from which cases fill the slots. RM carries over the reference's workload and allows the candidate to improve this second component. The exact cardinality–ranking decomposition, its population counterpart, and the behavior of the transported count under label shift are derived in Appendices D.1–D.2. The reference count is an existing operating point, not an estimate of the F-measure-optimal count. Our experiments use $\beta = 1$ and macro averaging.

4.2 When the evidence supports replacement

Exact count matching holds on the matching window. After the candidate cutoff is frozen, the two policies may select different fractions of fresh traffic. Certification therefore accounts for both the change in positive capture, $\Delta_g = g_S - g_R$, and the change in selection rate, $\Delta_\mu = \mu_S - \mu_R$. For $q > 0$, cross-multiplying the population F_β expressions from Section 3 gives

$$F_{\beta,S} > F_{\beta,R} \quad \Longleftrightarrow \quad (\beta^2 q + \mu_R)\Delta_g - g_R\Delta_\mu > 0. \tag{4}$$

A supported positive capture gain and a margin $\mu_R\Delta_g - g_R \max\{0, \Delta_\mu\} > 0$ are sufficient for improvement for every $\beta > 0$. This condition does not require estimating the target positive prevalence q.

The observations in Section 3 estimate these two changes on paired cases: $\Delta_g = u_+ - v_+$ uses held-out historical positives under capture-rate stability, and $\Delta_\mu = p_{10} - p_{01}$ uses independent unlabeled target traffic. Write underlines and overlines for one-sided confidence bounds. The direct-cell construction sets $\underline{\Delta}_g = \underline{u}_+ - \overline{v}_+$ and $\overline{\Delta}_\mu = \overline{p}_{10} - \underline{p}_{01}$, then computes

$$\Gamma_{\text{pair}} = \underline{\mu}_R\underline{\Delta}_g - \overline{g}_R \max\{0, \overline{\Delta}_\mu\}. \tag{5}$$

The first term credits the additional positives supported by historical evidence. The second charges for a possible increase in review volume on current traffic. C-RM deploys the candidate only when the first exceeds the second; otherwise it retains the reference.

Theorem 1 (paired certificate for decision transfer). Fix R and S using data independent of certification. Suppose $q > 0$ and $P(H = 1 \mid Y = 1) = Q(H = 1 \mid Y = 1)$ for $H \in \{R, S\}$. On independent source-positive and unlabeled target certification samples, construct the six one-sided Clopper–Pearson bounds $\underline{u}_+, \overline{v}_+, \overline{g}_R, \underline{\mu}_R, \overline{p}_{10}, \underline{p}_{01}$, each with error budget $\delta/6$. Then, conditional on the independent matching and policy-selection data,

$$\Pr(\Gamma_{\text{pair}} > 0,\ \exists \beta > 0 :\ F_{\beta,S}(Q) \le F_{\beta,R}(Q)) \le \delta. \tag{6}$$

This is a per-comparison statement about population performance: it bounds the joint event of approval and non-improvement, rather than error conditional on approval or the realized outcome of one finite rollout. The same approval covers all $\beta > 0$. Multiple tested policies use a family-wise budget; independent source selection may first retain one candidate per class. Appendices D.3–D.4 give proofs, the multi-comparison guarantee, and the additional margin needed for a fresh-batch guarantee.

4.3 Using disagreements more efficiently

The direct-cell test bounds positives gained and lost separately, although both counts arise from the same set of changed decisions. When these events are rare, subtracting two separately widened intervals can obscure a useful net gain. We instead separate the frequency of a change from its direction.

On source positives, let $d_+ = P(S \neq R \mid Y = 1)$ be the disagreement probability and $\theta_+ = P(S = 1, R = 0 \mid S \neq R, Y = 1)$ the candidate-only share among disagreements. Define $d_t = Q(S \neq R)$ and $\theta_t = Q(S = 1, R = 0 \mid S \neq R)$ on unlabeled target traffic. Then

$$\Delta_g = d_+(2\theta_+ - 1), \qquad \Delta_\mu = d_t(2\theta_t - 1). \tag{7}$$

The source factors distinguish how many positive cases change hands from whether those changes favor the candidate. The target factors measure the corresponding change in workload. We place exact-binomial bounds on each frequency and on each direction conditional on the observed disagreement count, then insert the resulting contrast bounds into the same approval margin.

Proposition 1 (disagreement-factorized paired gate). Under the conditions of Theorem 1, the factorized bounds in Appendix D.3.1, Equation 30, use the same samples and six allocations of $\delta/6$. Approving when $\Gamma_{\mathrm{fac}} > 0$ satisfies the erroneous-approval bound in Equation 6 for every $\beta > 0$ simultaneously.

Factorization uses the same supervision and risk budget more efficiently (Section 5.4). Historical positives resolve the capture difference, while unlabeled target cases resolve the workload difference.

Scope under changing traffic. Certification uses an explicit stability or bounded-drift assumption. Changes to the positive population require an externally justified allowance that reduces the certified capture gain (Appendix D.4). Unlabeled traffic alone cannot validate stability: identical observed decisions can coexist with opposite upgrade outcomes (Appendix D.5). The drift extension makes the assumed tolerance and remaining certification margin explicit.

5 Experiments

5.1 Controlled prior shift

We first isolate the count-selection component using 30 rare RCV1 labels (source prevalence approximately 0.2–0.9%). Logistic regression supplies the reference and a rank-averaged ensemble supplies the final ranking. Thresholds use 60,000 labeled validation examples. Each one-vs-rest class is independently resampled into a 20,000-example target batch with a prior multiplier in $\{1, 1.5, 2, 3, 5, 8\}$; macro-F1 averages class-wise F1 over 30 classes and 10 resampling seeds. This construction isolates prior shift but does not preserve a common multilabel target bag. Table 1 holds the receiving ranker fixed and compares four ways of choosing its selection count. RM transfers the reference count; the receiver's source-validation threshold, MLLS, and calibrated BBSE can select different counts. This is a comparison of count-selection rules, rather than an equal-workload comparison. A separate fixed-cardinality ablation and a target-label oracle are reported in Appendix C.

Table 1: RCV1 macro-F1 over rare classes. Receiver source threshold is calibrated separately for the ensemble on labeled source validation data.

	Target-prior multiplier				
Method	1×	2×	3×	5×	8×
Receiver source threshold	.5386	.6106	.6399	.6656	.6828
Rate matching	.5373	.6110	.6417	.6675	.6844
MLLS	.5309	.6095	.6521	.7070	.7556
Calibrated BBSE	.5280	.6098	.6542	.7092	.7567

RM is strongest near the calibrated source regime: it exceeds calibrated BBSE by .0094 at $1\times$ (95% class-bootstrap CI $[.0030, .0156]$), while BBSE and MLLS dominate at $5\times$–$8\times$. At exactly the reference count, changing only the ranker improves macro-F1 by .0070 at $1\times$ and .0044 at $8\times$, isolating ranking from action quantity. Full validation-size and fixed-cardinality analyses are in Appendix C.

5.2 Cross-model transfer with deep representations

We next change architecture and score scale at a fixed reference count. GoEmotions (Demszky et al., 2020) uses a TF–IDF reference and source-fine-tuned text encoders; CIFAR-100 (Krizhevsky, 2009) uses frozen ResNet-18 and ConvNeXt-Tiny representations (He et al., 2016; Liu et al., 2022). COCO covers 117,266 images and three candidate architectures. All class and checkpoint choices use source data. For every class and target resample, Table 2 compares rankers at exactly the reference's selection count. Training and split details appear in Appendix E.

Table 2: Cross-model macro-F1 at the reference-derived count. Values average three independent receiver/head initializations and ten target resampling seeds. Shading marks the highest mean within each dataset at all displayed settings.

		Target-prior multiplier				
Dataset	Ranking	1×	2×	3×	5×	8×
GoEmotions	TF–IDF reference	.4199	.4896	.5222	.5563	.5785
	weighted RoBERTa	.4751	.5453	.5751	.6041	.6153
	Qwen3-0.6B	.4821	.5514	.5801	.6072	.6181
CIFAR-100	ResNet-18 reference	.4798	.5497	.5780	.6057	.6231
	ConvNeXt-Tiny	.6336	.6817	.6871	.6846	.6778
COCO-117k	ResNet-50 reference	.6309	.6900	.7107	.7325	.7433
	DINOv2-base	.7129	.7587	.7675	.7717	.7697
	SigLIP2-base	.7533	.7867	.7882	.7849	.7771
	ConvNeXt-Tiny	.6755	.7289	.7439	.7574	.7602

At the transported count, the strongest GoEmotions, CIFAR-100, and COCO receivers gain .0622, .1538, and .1224 macro-F1 at $1\times$; gains remain positive at $8\times$. Negative controls using unweighted or generic frozen receivers lose F1, showing that architecture scale alone is insufficient. Five COCO partitions, common multilabel bags, MLP heads, and encoder fine-tuning preserve positive transfer (Appendix C).

ModernBERT-base (Warner et al., 2024) sharpens Proposition D.3: across three seeds its macro-AUC (.9300) exceeds weighted RoBERTa's (.9265), yet its fixed-count F1 is lower at every tested depth (Figure 2). Both beat the reference. Source data fix all policies; target labels only audit the curves.

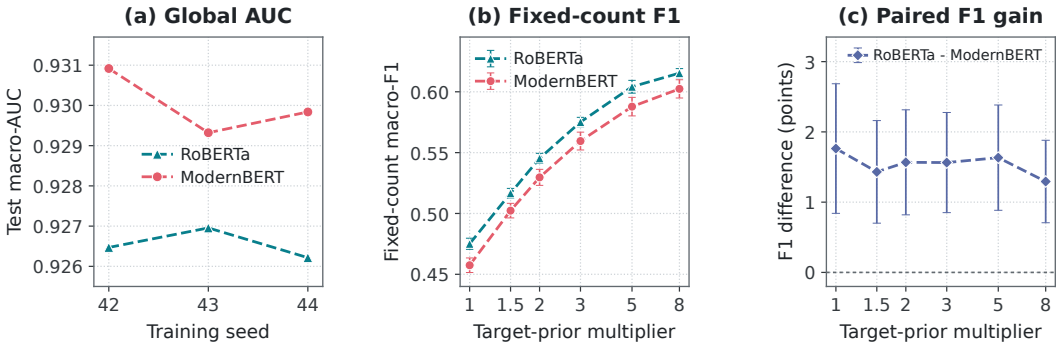

Figure 2: Higher global AUC need not yield better decisions at a matched workload. (a) ModernBERT has higher test macro-AUC in each training seed. (b) RoBERTa has higher macro-F1 at every TF–IDF-reference-derived count. (c) Paired RoBERTa-minus-ModernBERT gains are 1.29–1.76 F1 points (one point is .01 F1). Panels (b,c) use log-scaled prior multipliers and show means with one sample standard deviation over three training seeds. Target labels audit these source-frozen policies; they are not used for deployment selection.

5.3 Certified model replacement

Table 3: Strict GoEmotions model-upgrade audit. The disjoint source selector prefers the candidate in both rows; C-RM assesses each local replacement. Each row covers three receiver seeds, 25 classes, seven prior multipliers, and 50 target draws per seed. Approved switches are conservative/fully paired counts; losses are population/finite-batch counts among approved actions.

Upgrade	Naive target ΔF1	Harmful comparisons	Approved C-RM switches	Losses pop./batch
TF–IDF $\rightarrow$ Qwen	+.0444	3,578/26,250	225/834	0/0
RoBERTa $\rightarrow$ ModernBERT	$-.0115$	17,649/26,250	0/0	–

–: no approved replacement; the reference is retained.

The source selector prefers ModernBERT in every seed, yet its target F1 changes are $-.0132/-.0128/-.0084$. C-RM retains RoBERTa while admitting supported upgrades from the weaker TF–IDF reference. Global model selection proposes a candidate; local evidence determines which decisions to replace. Fully paired C-RM approves 3.18% of TF–IDF-to-Qwen comparisons: mean deployed population F1 rises from .5282 to .5311, versus .5726 for unconditional replacement. The gate realizes part of the average gain while avoiding all observed harmful approvals in this audit.

5.4 Certification power and robustness

We evaluate Theorem 1 on all 117,266 labeled COCO images using 30 fixed classes, three independent head pairs, seven prior multipliers, and 100 i.i.d. target draws of size 5,000. Threshold selection, source certification, unlabeled rate certification, and evaluation use disjoint samples. Table 4 tests the power gain from disagreement factorization. The direct-cell and factorized gates use the same receiver, samples, six exact-binomial bounds, and nominal risk.

Table 4: Disagreement factorization increases certificate power at $\delta = .05$. Each source split contains 31,500 pre-specified comparisons; counts use identical comparisons and supervision within a row. Target labels are used only for retrospective loss auditing.

	Approved switches			
Source split	Direct-cell	Factorized	Power gain	Population losses
1	153	616	$4.0\times$	0
2	190	846	$4.5\times$	0
3	95	439	$4.6\times$	0

Nominal-risk calibration. To measure approval error and power, we simulate fixed policy pairs whose source-positive disagreement orientation sets the true TPR and F1 gap, while the unlabeled target joint-decision table is held fixed and all C-RM assumptions are satisfied. Table 5 reports 50,000 independent repetitions per row at $\delta = .05$.

Table 5: Controlled calibration and power at 5,000 source positives and 20,000 unlabeled target examples. Entries are approval percentages with exact 95% binomial intervals. Approvals in non-positive rows are errors; positive rows measure power.

	Direct-cell gate		Factorized gate	
True F1 gain	Approval (%)	95% interval	Approval (%)	95% interval
$-.0067$	.000	$[.000, .007]$	.000	$[.000, .007]$
.0000	.000	$[.000, .007]$	.008	$[.002, .020]$
$+.0067$	.010	$[.003, .023]$	.504	$[.444, .570]$
$+.0333$	40.432	$[40.001, 40.864]$	72.782	$[72.390, 73.172]$

Across the full 3×3 evidence grid, the largest harmful approval rate is .008% and the largest null approval rate is .044%, both far below the 5% per-comparison budget. Factorization increases power as positive and unlabeled support grow while keeping these empirical errors below the budget.

All approved COCO switches have nonnegative population F_β gain for $\beta \in \{.5, 1, 2\}$. The weaker ConvNeXt receiver yields less certificate support and the stronger SigLIP2 yields more, linking approval coverage to candidate-specific evidence.

Family-wise risk $.05/30$ retains 68 factorized, non-losing upgrades. Coverage also increases with unlabeled evidence in the tested settings; detailed candidate-selection and evidence-scaling results are in Appendix C.

Under natural CivilComments shift, source-fine-tuned ModernBERT improves global F1 from .5819 to .6553; C-RM approves 108/150 non-losing global comparisons while retaining the reference on low-support identity-group comparisons. This replicates the same evidence-adaptive behavior as the controlled COCO study under a natural text-distribution shift. Table 7 compares three source-fine-tuned candidates on the same global and subgroup audit.

6 Conclusion

RM separates reference-derived action volume from candidate ranking. C-RM evaluates the resulting fixed-policy replacement using historical positives and unlabeled target traffic. The paired analysis links this evidence to population F_β improvement, while disagreement factorization increases certification power at the same nominal risk level. In the GoEmotions audit, C-RM admits a subset of beneficial Qwen replacements and retains RoBERTa against a source-better but target-worse ModernBERT. This procedure evaluates model upgrades at the deployment operating point before target outcomes arrive.

AI USE STATEMENT

In this study, we used generative AI tools for literature searches and translation, grammatical correction and language polishing of the manuscript, and code refactoring to improve readability. We have reviewed all AI-generated paper summaries for accuracy. The code refactoring did not alter functionality. We take responsibility for the final content of this work.

REFERENCES

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 222–232, 2020.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021. URL https://arxiv.org/abs/2110.01052.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022. URL https://arxiv.org/abs/2208.02814.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*, 2021. URL https://arxiv.org/abs/2101.02703.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Bastien Dussap, Gilles Blanchard, and Badr-Eddine Chérief-Abdellatif. Label shift quantification with robustness guarantees via distribution feature matching. *arXiv preprint arXiv:2306.04376*, 2023.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Sunay Joshi, Tao Wang, Hamed Hassani, and Edgar Dobriban. Risk-controlled post-processing of decision policies. *arXiv preprint arXiv:2605.06479*, 2026.

Kevin Kasa, Zhiyu Zhang, Heng Yang, and Graham W. Taylor. Adapting prediction sets to distribution shifts without labels. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1990–2010, 2025.

Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable learning with distribution shift. In *Proceedings of the Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 2887–2943, 2024.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Bert Earnshaw, Imran Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

Oluwasanmi Koyejo, Nagarajan Natarajan, Pradeep K. Ravikumar, and Inderjit S. Dhillon. Consistent binary classification with generalized performance metrics. *Advances in Neural Information Processing Systems*, 27, 2014.

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

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661. PMLR, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130, 2018.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Harikrishna Narasimhan, Purushottam Kar, and Prateek Jain. Optimizing non-decomposable performance measures: A tale of two classes. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 199–208, 2015.

Patrik Okanovic, Andreas Kirsch, Jannes Kasper, Torsten Hoefler, Andreas Krause, and Nezihe Merve Gürel. All models are wrong, some are useful: Model selection with limited labels. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2035–2043, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. A comparative evaluation of quantification methods. *Journal of Machine Learning Research*, 26(55):1–54, 2025.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6368–6377, 2020.

Dirk Tasche. Does quantification without adjustments work? *arXiv preprint arXiv:1602.08780*, 2016.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388. PMLR, 2015.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14299–14308, 2021.

A C-RM Algorithm and Deployment

Algorithm 1 Certified rate matching for one class

Require: Reference score r, threshold τ, receiver score s; disjoint source positives D_+^s and unlabeled target subsets $D_t^{\mathrm{match}}, D_t^{\mathrm{cert}}$; risk δ

1: $\widehat{k} \leftarrow \sum_{x \in D_t^{\mathrm{match}}} \mathbf{1}[r(x) \geq \tau]$
2: Choose t with pre-specified tie breaking so the receiver selects $\widehat{k}$ examples in D_t^{match}; freeze the tie rule and define $R = \mathbf{1}[r \geq \tau]$ and $S = \mathbf{1}[s \geq t]$
3: On D_+^s, count S-only, R-only, and reference-positive events; on D_t^{cert}, count both disagreement directions and the reference rate
4: Form six one-sided Clopper–Pearson bounds at risk $\delta/6$ and compute Γ_{pair} from Eq. (5)
5: **if** $\Gamma_{\mathrm{pair}} > 0$ **then**
6: **return** fixed receiver policy S for fresh rollout traffic
7: **else**
8: **return** complete reference policy R
9: **end if**

Table 6: C-RM deployment timeline. Matching, certification, and rollout use independent samples. Audit labels belong to the rollout cases and become available after deployment.

Stage	Evidence	Frozen operation	Meaning
Match	unlabeled target window	fit and freeze t at the reference count	same count, candidate-ranked cases
Certify	held-out source positives and target traffic	estimate paired changes and compute the margin	approve population improvement at comparison risk δ
Roll out	fresh target traffic	apply fixed S if approved; otherwise R	execute the selected policy
Audit	delayed labels for rollout cases	measure realized F_β	assess the deployed decisions retrospectively

B Count-Transport Diagnostics

Proposition B.1 (adjusted-count variance). The population identity is $q = (\mu_Q - b)/(a - b)$. For independent estimates from n target examples and an m-example labeled validation set, the delta method gives

$$\mathrm{Var}(\widehat{q}_{AC}) \approx \frac{1}{J^2}\left[\frac{\mu_Q(1-\mu_Q)}{n} + \frac{q^2 a(1-a)}{mp} + \frac{(1-q)^2 b(1-b)}{m(1-p)}\right]. \tag{8}$$

The TPR estimate uses only about mp positives, creating a rare-class term proportional to $1/(mpJ^2)$. This explains why stable but biased RM can win at small shifts, while adjusted methods win after RM's bias dominates their variance.

First-order crossover diagnostic. For a realized batch, the perfect-ranker cardinality fraction is N_+/n. To obtain a population diagnostic before labels are observed, we compare normalized RM and adjusted count with its limit q. Their first-order mean-squared count errors relative to this

population benchmark are

$$\mathrm{MSE}_{RM} = (\mu_Q - q)^2 + \frac{\mu_Q(1-\mu_Q)}{n}, \tag{9}$$

$$\mathrm{MSE}_{AC} \approx \frac{1}{J^2}\left[\frac{\mu_Q(1-\mu_Q)}{n} + \frac{q^2 a(1-a)}{mp} + \frac{(1-q)^2 b(1-b)}{m(1-p)}\right]. \tag{10}$$

Under the first-order approximation, RM has the lower population count-risk proxy whenever the first expression is smaller than the second. The comparison exposes the regime boundary: small mp or small $|J|$ penalizes inversion, whereas large validation sets reduce adjusted-count variance and leave RM's policy mismatch $(\mu_Q - q)^2$. Corollary D.5 translates that mismatch into the actual F_β target while preserving local ranking regret. The MSE comparison uses the population benchmark q; the finite-batch optimum N_+/n additionally fluctuates around it.

Approximate label shift. If target conditional rates are $a_Q = a + \delta_a$ and $b_Q = b + \delta_b$, then

$$\mu_Q = b + Jq + q\delta_a + (1-q)\delta_b. \tag{11}$$

Thus the label-shift prediction of the transported rate incurs at most $q|\delta_a| + (1-q)|\delta_b|$ error. RM still exactly counts the reference model's target predictions, but their interpretation as a source policy adapting with slope J degrades; adjusted-count inversion is also misspecified because it uses source conditional rates.

Derivation of adjusted-count variance. Apply the multivariate delta method to $g(\mu, a, b) = (\mu - b)/(a - b)$. At the population identity, $\partial_\mu g = 1/J$, $\partial_a g = -q/J$, and $\partial_b g = -(1-q)/J$. Substituting independent Bernoulli variances for target decisions, validation positives, and validation negatives yields the stated expression. This approximation conditions on a fixed threshold and omits covariance from shared threshold estimation.

Derivation of the crossover diagnostic. RM's normalized count is a Bernoulli sample mean with expectation μ_Q, so its MSE relative to q is squared bias $(\mu_Q - q)^2$ plus variance $\mu_Q(1-\mu_Q)/n$. Adjusted count is first-order unbiased under the assumed model, so Proposition B.1 supplies its approximate MSE. Comparing the two expressions gives the stated condition.

Derivation under approximate label shift. Substitute $Q(A \mid Y = 1) = a + \delta_a$ and $Q(A \mid Y = 0) = b + \delta_b$ into the law of total probability and apply the triangle inequality.

C Extended Empirical Results

Interpreting deployment results. RM applies classical CC to reference-derived workload. Its utility depends on that count and the candidate's local ranking. Approval coverage and realized gain accompany loss counts throughout these audits. More source data tightens sampling intervals; positive-conditional drift is treated through the declared allowances in Proposition D.8. As in Section 4, the certificate evaluates a frozen policy, not exact top-k re-selection on every future batch. Corollary D.7 controls multiple updates and Proposition D.6 provides the prevalence floor and margin for a fresh-batch guarantee. Finite-batch losses elsewhere are retrospective audits.

Count selection and ranking controls. RM is competitive with the receiver's source-validation threshold. The two are close near $1\times$–$2\times$ in this same-family percentile-rank experiment, but the latter uses receiver-specific labeled validation calibration and is not old numerical-threshold reuse. Against calibrated BBSE, RM's paired class-level advantage is $.0094$ at $1\times$ (95% bootstrap CI $[.0030, .0156]$), becomes inconclusive by $1.5\times$–$2\times$, and reverses at $3\times$; BBSE and MLLS are clearly stronger at $5\times$ and $8\times$. Shrinking validation size from 60,000 to 5,000 increases RM's near-no-shift advantage over calibrated BBSE from about .009 to .039, consistent with the $1/(mpJ^2)$ term.

A target-label oracle freely optimizes action volume and scores .5659/.6385/.6799/.7310/.7748 over the five reported multipliers. This is the unconstrained F1 ceiling; RM instead preserves the reference volume.

We directly isolate the ranking term in a separate fixed-cardinality ablation. For each target resample, both the logistic reference and the rank-averaged ensemble select exactly the number of examples produced by thresholding the reference. Replacing only the ranking raises macro-F1 by .0070, .0068, .0055, and .0044 at $1\times$, $2\times$, $5\times$, and $8\times$, respectively (30 classes and 10 target seeds). These fixed-count gains isolate the contribution of ranking quality.

The strongest GoEmotions receiver gains $+.0622$ at $1\times$ and $+.0396$ at $8\times$; on CIFAR-100 the gains are $+.1538$ and $+.0547$, respectively, with sample standard deviation below .0010 for both CIFAR gains across three head initializations. On all 36 available COCO (Lin et al., 2014) labeled-data shards (117,266 images), replacing ResNet-50 with DINOv2-base (Oquab et al., 2024) raises validation AUC from .9789 to .9825 and fixed-count F1 by .0820 at $1\times$ and .0264 at $8\times$. SigLIP2-base (Tschannen et al., 2025) raises validation AUC to .9899 and gives the largest gains, .1224 and .0338; ConvNeXt-Tiny gives smaller but still positive gains of .0446 and .0169. Useful count transport spans three upgraded architectures. Receiver training also changes local ranking: an unweighted RoBERTa has test macro-AUC .8375 and loses .1224 macro-F1 at the same $1\times$ count; frozen MiniLM and generic Qwen3-0.6B mean-pooled representations also transfer negatively. Positive-weight caps 6, 10, and 20 all reverse the sign, showing that receiver training must improve rare-class ranking near the deployed depth. Across the three weighted RoBERTa runs, source-validation gain at the reference depth has per-class Spearman correlation .416–.647 with target fixed-count gain at $1\times$, compared with only .090–.265 for global AUC gain.

Across the same three seeds, ModernBERT reaches test macro-AUC $.9300 \pm .0008$, above weighted RoBERTa's $.9265 \pm .0004$, but fixed-count F1 is $.4575 \pm .0060$ versus $.4751 \pm .0046$ at $1\times$ and $.6024 \pm .0075$ versus $.6153 \pm .0037$ at $8\times$.

Source-fine-tuning Qwen3-0.6B rather than mean-pooling its generic representation reaches test macro-AUC $.9391 \pm .0013$ across three seeds and the strongest mean fixed-count F1 at every reported depth. In the strict audit of Table 3, model selection and certification use disjoint validation subsets. Qwen raises selector macro-AUC from .8946 to .9410, yet a naive class-local rollout loses population F1 in 3,578 of 26,250 comparisons. Conservative/fully paired C-RM accepts 225/834 upgrades with no population or finite-batch loss. Together with CivilComments, this recent decoder check replicates across two datasets and six independent training runs.

For the GoEmotions quantification control, a fixed three-seed RoBERTa rank ensemble is paired with PACC, MLLS, source-tuned RLLS (Azizzadenesheli et al., 2019), and RBF distribution feature matching (DFM) (Dussap et al., 2023). At $1\times$, RM scores .4830 versus .4736, .4745, .4569, and .4541, respectively. At $3\times$, PACC reaches .5918 versus RM's .5801; at $8\times$, MLLS reaches .7147 versus RM's .6186. These quantifiers exhibit the same regime change: reference-derived decision transport performs best near the calibrated source regime, while explicit prior estimation wins under large controlled shifts. The .4821 entry in Table 2 uses source-fine-tuned Qwen; the .4830 RM quantifier control uses the three-seed RoBERTa rank ensemble.

The full COCO experiment also includes a common-bag stress test that preserves all labels attached to each sampled image. Enriching one source-selected anchor label creates one shared target bag for all 30 evaluated labels, rather than a separate resample per label. DINOv2 improves fixed-count macro-F1 by .0792 at $1\times$ and .0795 at the largest enrichment, showing that the gain survives multilabel co-occurrence. Across five full-data partitions, per-class gains are $.0830 \pm .0032$, $.0529 \pm .0037$, and $.0240 \pm .0020$ at $1\times$, $3\times$, and $8\times$; the corresponding shared-bag gains are $.0822 \pm .0032$, $.0829 \pm .0029$, and $.0817 \pm .0030$. On all 117,266 images, replacing the linear heads with two-layer 512-dimensional MLPs also preserves positive transfer: the per-class gains are .0514, .0290, and .0122, while the shared-bag gains are .0473, .0469, and .0431. Positive transfer therefore persists across these partitions and head architectures. Fine-tuning both encoders on the source split before extracting representations gives the same conclusion across three encoder training seeds: per-class gains are $.0551 \pm .0125$, $.0352 \pm .0088$, and $.0158 \pm .0047$ at $1\times$, $3\times$, and $8\times$, and every run is positive. The corresponding shared-bag gains are .0557, .0530, and .0531.

Under natural RCV1 temporal drift, macro-F1 is .5274 for the receiver's source-validation threshold, .5289 for RM, .4743 for MLLS, .4589 for calibrated BBSE, and .5300 for the RM/BBSE selector, indicating that prior inversion is brittle when pure label shift is violated.

In a separate RoBERTa-to-Qwen upgrade audit, Qwen wins the disjoint source selector for every seed. Target mean F1 changes are $-.0023$, $+.0152$, and $-.0015$ for seeds 42, 43, and 44, respectively. The sign varies across training runs even after source selection.

With DINOv2 and ConvNeXt available together, the independent source selector chooses them for 66 and 24 of 90 class/head policies. Conservative C-RM then makes 165 non-losing switches, essentially retaining the 166 switches of the single-DINOv2 run; full pairing makes 353. In contrast, certifying both candidates at $\delta/2$ abstains throughout. Independent preselection therefore preserves statistical power using source-only candidate choice. All conservative switches in this audit come from DINOv2. Across three source splits and the same first 50 target draws, independent selection yields 65/70/53 conservative and 154/192/97 fully paired switches; all are non-losing on both evaluation objects. DINOv2/ConvNeXt are selected for 206/64 of 270 class/head policies, so the candidate choice responds to source evidence rather than being fixed by architecture name. Across the same three splits, allocating 1/2.5/5% of source data to the disjoint selector gives 206/246/188 conservative and 441/550/443 fully paired switches, all non-losing; extra selector data does not monotonically increase coverage because it changes both policy choice and the remaining evidence.

At the primary $\delta = .05$, five additional source splits make 808 certified switches with no population or finite-batch loss. Allocating family-wise risk $\delta_{\mathrm{F}} = .05$ over 30 classes retains 68 factorized, non-losing upgrades, demonstrating that the disagreement parameterization preserves power under a deployment-level multiplicity budget. Increasing the unlabeled certificate subset from 500 to 5,000 examples raises conservative coverage from 7 to 166 switches, while population loss remains zero and finite-batch losses fall from two to zero. These results isolate the two intended benefits of the factorized certificate: greater power from the same supervision and predictable improvement as target evidence accumulates.

Table 7: Natural-shift audit on CivilComments, with a TF–IDF reference (mean global F1 .5819). Each candidate uses three training seeds and 50 target partitions per seed. F1 is the candidate's unconditional mean on the global partitions; losses count approved global replacements with lower partition F1. Subgroup columns cover eight identity groups. These are retrospective shift diagnostics with unverified capture-rate stability.

	Global partitions			Identity groups
Candidate	Mean F1	Approved / tested	Losses	Approved / tested
ModernBERT	.6553	108/150	0	0/1,200
RoBERTa	.6567	90/150	0	0/1,200
Qwen3-0.6B	.6595	141/150	0	0/1,200

Natural subpopulation shift. We then test natural subpopulation shift on WILDS CivilComments (Koh et al., 2021), splitting validation data between selection and disjoint source-positive certification. A frozen ModernBERT head is a useful negative control: at the reference-derived count its F1 is .4886 versus .5695 for the TF–IDF reference. Forced replacement would lower F1 in 399/400 identity-group comparisons, while C-RM rejects all 400. After source-only ModernBERT fine-tuning, global test-batch RM raises F1 from .5819 to $.6553 \pm .0020$ across three seeds. The fully paired gate accepts 34/31/43 of 50 target partitions, respectively, with zero losses. In contrast, it rejects all 1,200 comparisons over eight identity groups because group-specific source-positive evidence is insufficient: the supported global upgrade proceeds while these groups retain the old policy. Three RoBERTa seeds likewise reach fixed-count F1 $.6567 \pm .0003$, and C-RM accepts 90/150 global comparisons with zero loss and again makes no identity-group switch. These audits measure behavior under natural shift, where positive-conditional stability is unverified.

Stronger-supervision diagnostics. On the Camelyon17 hospital shift, a random budget of 20 or 50 labeled target examples retains no proposed upgrade. Table 8 reports larger budgets and a privileged positive-only sample. Allocating .05 family-wise over these comparisons abstains throughout. In this audit, small target samples do not resolve the positive-conditional mismatch. Separately, a target-calibrated risk-controlled post-processing diagnostic uses 5,000 labeled target pairs per comparison and controls a pointwise violation objective. Its supervision and objective differ from target-unlabeled, equal-volume F_β replacement.

Table 8: Target-label budget diagnostic on the Camelyon17 hospital shift. Each row evaluates 1,000 proposed comparisons. The positive-only sample is privileged: its budget excludes the labels needed to find those positives. Losses count approved replacements with lower population F1.

Target-label access	Budget	Approved switches	Population losses
Random examples	20	0	0
Random examples	50	0	0
Random examples	100	6	3
Random examples	200	46	13
Positives only (privileged)	200	186	34

D Additional Theory and Proofs

This appendix gives the supporting decision-depth analysis, transported-count identities, and extensions of the two main certification results. The assumptions and scope of each extension are stated with its result.

D.1 Selection count and local ranking

Let $N_+ > 0$ be the number of positives in a finite target batch, $TP_s(k)$ the positives selected by score s at depth k, and $s^\star$ a score that ranks every positive before every negative.

Proposition D.1 (exact cardinality–ranking decomposition). For a batch with $N_+ > 0$, any $\beta > 0$, score s, and cardinality $k \in \{0, \dots, n\}$,

$$1 - F_{\beta,s}(k) = \underbrace{1 - F_{\beta,s^\star}(k)}_{\mathcal{R}_{\mathrm{card}}(k)} + \underbrace{F_{\beta,s^\star}(k) - F_{\beta,s}(k)}_{\mathcal{R}_{\mathrm{rank}}(s;k)}, \tag{12}$$

$$\mathcal{R}_{\mathrm{card}}(k) = \begin{cases} \dfrac{\beta^2(N_+ - k)}{\beta^2 N_+ + k}, & k \le N_+, \\ \dfrac{k - N_+}{\beta^2 N_+ + k}, & k > N_+, \end{cases} \tag{13}$$

$$\mathcal{R}_{\mathrm{rank}}(s;k) = \frac{(1+\beta^2)\{\min(k, N_+) - TP_s(k)\}}{\beta^2 N_+ + k}. \tag{14}$$

In particular, $\mathcal{R}_{\mathrm{card}}(k) \le |k - N_+|/N_+$ for $N_+ > 0$. The perfect target decision supplies the benchmark; an imperfect ranker's F-measure-optimal count can differ from N_+. In decision transfer, the reference sets the cardinality term and the receiver determines ranking error at the realized depth.

Proof. Expanding the definition of F_β with $FN = N_+ - TP_s(k)$ and $FP = k - TP_s(k)$ gives $F_{\beta,s}(k) = (1+\beta^2)TP_s(k)/(\beta^2 N_+ + k)$. A perfect ranking has $TP_{s^\star}(k) = \min(k, N_+)$. Adding and subtracting $F_{\beta,s^\star}(k)$ yields the decomposition. Substitution of $k \le N_+$ and $k > N_+$ gives the two cardinality expressions; subtracting the two F-measures gives the ranking term. For $k \le N_+$, the cardinality denominator is at least $\beta^2 N_+$; for $k > N_+$, it exceeds N_+, proving the stated bound.

Local ranking at fixed depth. At a fixed transported k, replacing score s_1 by s_2 does not reduce F_β if and only if $TP_{s_2}(k) \ge TP_{s_1}(k)$, and improves it strictly if and only if the inequality is strict. This is a local precision-at-depth condition; larger global AUC alone is insufficient.

Proposition D.2 (population marginal-depth condition). Let the top-u decision under score s select target mass $u \in (0,1)$, using randomized boundary ties when necessary, and write $T_s(u) = Q(Y = 1, S_s(u) = 1)$ for its selected positive mass. Assume T_s is differentiable at u and define the marginal precision $\eta_s(u) = T_s'(u)$. Then

$$\operatorname{sign} F'_{\beta,s}(u) = \operatorname{sign}\left\{\eta_s(u) - \frac{F_{\beta,s}(u)}{1+\beta^2}\right\}. \tag{15}$$

If T_s is twice differentiable and η_s is nonincreasing, the population F_β curve is unimodal; any interior optimum $u_s^\star$ satisfies $\eta_s(u_s^\star) = F_{\beta,s}(u_s^\star)/(1+\beta^2)$. This condition is in quantile space and needs no numerical score calibration, but global optimality does require monotone target marginal precision. Moreover, if $F_{\beta,s}$ is twice differentiable near a strict interior local optimum $u_s^\star$ and $|F''_{\beta,s}(v)| \le L$ there, then for u in that neighborhood the local depth penalty obeys

$$0 \le F_{\beta,s}(u_s^\star) - F_{\beta,s}(u) \le \frac{L}{2}(u - u_s^\star)^2. \tag{16}$$

The quadratic statement is local and smooth; it does not replace the exact piecewise-rational cardinality term in Proposition D.1.

Proof. In population form, $F_{\beta,s}(u) = (1+\beta^2)T_s(u)/(\beta^2 q + u)$. Differentiation gives

$$F'_{\beta,s}(u) = \frac{(1+\beta^2)\{\eta_s(u)(\beta^2 q + u) - T_s(u)\}}{(\beta^2 q + u)^2}, \tag{17}$$

which proves Equation 15. If η_s is nonincreasing, then $h(u) = \eta_s(u)(\beta^2 q + u) - T_s(u)$ is nonincreasing wherever differentiable, because $h'(u) = \eta'_s(u)(\beta^2 q + u) \le 0$. Hence the derivative changes sign at most once, proving unimodality and the interior equality. Finally, Taylor's theorem around an interior optimum with $F'_{\beta,s}(u_s^\star) = 0$ gives $F_{\beta,s}(u) = F_{\beta,s}(u_s^\star) + \frac{1}{2}F''_{\beta,s}(\xi)(u - u_s^\star)^2$ for an intermediate ξ; the curvature bound proves Equation 16.

Proposition D.3 (global AUC can be arbitrarily misleading at fixed depth). For every fixed integer $k \ge 1$ and every $\varepsilon > 0$, there is a finite binary-ranking problem and scores s_1, s_2 such that

$$\mathrm{AUC}(s_2) - \mathrm{AUC}(s_1) > 1 - \varepsilon, \qquad TP_{s_2}(k) < TP_{s_1}(k). \tag{18}$$

Thus even an almost maximal AUC advantage can select fewer positives at a pre-specified deployment depth. The ModernBERT/RoBERTa crossover in Section 5.2 is a real-data instance of this distinction between global ranking and local decision quality.

Proof. Choose integers $P > \max\{k, 2k/\varepsilon\}$ positives and $M > \max\{1, 2/\varepsilon\}$ negatives. Let s_1 rank k positives first, then all M negatives, then the remaining $P - k$ positives. Thus $TP_{s_1}(k) = k$ and $\mathrm{AUC}(s_1) = k/P$. Let s_2 rank $k-1$ positives first, then one negative, then the remaining $P - k + 1$ positives, and finally the other $M - 1$ negatives. Now $TP_{s_2}(k) = k - 1$, and only the first negative creates positive–negative inversions, so

$$\mathrm{AUC}(s_2) = 1 - \frac{P - k + 1}{PM}. \tag{19}$$

Therefore $1 - \{\mathrm{AUC}(s_2) - \mathrm{AUC}(s_1)\} \le 1/M + k/P < \varepsilon$, proving the claim.

D.2 Transported-count behavior

Now let $a = P(A \mid Y = 1) = Q(A \mid Y = 1)$, $b = P(A \mid Y = 0) = Q(A \mid Y = 0)$, and $J = a - b$ under label shift.

Lemma D.4 (transported rate). The source and target positive-decision rates satisfy

$$\mu_P = b + Jp, \qquad \mu_Q = b + Jq = \mu_P + J(q - p). \tag{20}$$

Conditional on the frozen reference and threshold, $\mathbb{E}[\widehat{k}_{RM}/n] = \mu_Q$ and $\mathrm{Var}(\widehat{k}_{RM}/n) = \mu_Q(1 - \mu_Q)/n$. RM is therefore unbiased for the reference's target *prediction rate*, not for target prevalence or the F1-optimal rate. The slope J quantifies partial adaptation to prior change.

Proof. By total probability, $Q(A) = Q(A \mid Y = 1)q + Q(A \mid Y = 0)(1 - q) = b + Jq$. The source identity is analogous. The target indicators are i.i.d. Bernoulli(μ_Q) under the batch-sampling assumption.

Corollary D.5 (end-to-end population RM regret). For a target selection rate u, define $T_s(u) = Q(Y = 1, S_s(u) = 1)$ and let a perfect target ranking select positive mass $\min(u, q)$. The normalized analogue of Proposition D.1 is

$$1 - F_{\beta,s}(u) = \mathcal{R}^Q_{\mathrm{card}}(u) + \mathcal{R}^Q_{\mathrm{rank}}(s; u), \tag{21}$$

$$\mathcal{R}^Q_{\mathrm{card}}(u) = \begin{cases} \dfrac{\beta^2(q-u)}{\beta^2 q + u}, & u \le q, \\ \dfrac{u-q}{\beta^2 q + u}, & u > q, \end{cases} \tag{22}$$

$$\mathcal{R}^Q_{\mathrm{rank}}(s; u) = \frac{(1+\beta^2)\{\min(u, q) - T_s(u)\}}{\beta^2 q + u}. \tag{23}$$

Under label shift, the transported reference rate $u_R = \mu_Q$ satisfies the exact mismatch identity

$$u_R - q = (\mu_P - p) + (J - 1)(q - p). \tag{24}$$

Consequently, for $q > 0$,

$$1 - F_{\beta,s}(u_R) \le \frac{|(\mu_P - p) + (J - 1)(q - p)|}{q} + \mathcal{R}^Q_{\mathrm{rank}}(s; u_R). \tag{25}$$

If the source action rate is cardinality-calibrated ($\mu_P = p$), $J = 1$ tracks prior change exactly, whereas $J \approx 0$ transports an almost constant rate and incurs mismatch approximately $|q - p|$. This is an F_β regret statement, unlike the count-MSE diagnostic in Appendix B; it retains receiver ranking error explicitly. The $1/q$ term can make this upper bound exceed one for rare labels. Equation 21 gives the exact decomposition, and reported F_β values use exact confusion counts. Deployment uses the paired margin of Theorem 1.

Proof. Normalize Proposition D.1 by the target population size and replace k/N, N_+/N, and $TP_s(k)/N$ by u, q, and $T_s(u)$, respectively. This gives Equation 21. Subtracting the source identity $\mu_P = b + Jp$ from the target identity $\mu_Q = b + Jq$ and then subtracting $q - p$ gives Equation 24. The bound follows from $\mathcal{R}^Q_{\mathrm{card}}(u) \le |u - q|/q$ for $q > 0$.

D.3 Proofs of the certification guarantees

Proof of Theorem 1. For a policy H, write $g_H = Q(H = 1 \mid Y = 1)$ and $\mu_H = Q(H = 1)$. Direct substitution gives

$$F_{\beta,H} = \frac{(1+\beta^2) q g_H}{\beta^2 q + \mu_H}. \tag{26}$$

After cross multiplication, the sign of $F_{\beta,S} - F_{\beta,R}$ is the sign of

$$(\beta^2 q + \mu_R)(g_S - g_R) - g_R(\mu_S - \mu_R). \tag{27}$$

On the simultaneous confidence event, $g_S - g_R = u_+ - v_+ \ge \underline{u}_+ - \overline{v}_+ = \underline{\Delta}_g$, $g_R \le \overline{g}_R$, $\mu_R \ge \underline{\mu}_R$, and $\mu_S - \mu_R = p_{10} - p_{01} \le \overline{p}_{10} - \underline{p}_{01} = \overline{\Delta}_\mu$. If $\Gamma_{\mathrm{pair}} > 0$, then $\underline{\Delta}_g > 0$ and the displayed difference is positive after dropping the nonnegative $\beta^2 q(g_S - g_R)$ term. Each of the six constituent Clopper–Pearson bounds fails with probability at most $\delta/6$; a union bound proves joint coverage at least $1 - \delta$. The paired disagreement cells need not be independent for this union-bound argument. The coverage statement is conditional on the independent matching and threshold-selection sample, which makes the learned policies fixed; averaging over that split also gives the corresponding unconditional coverage.

D.3.1 Factorized confidence bounds

Construction. Let $d_+ = P(S \ne R \mid Y = 1)$ and $\theta_+ = P(S = 1, R = 0 \mid S \ne R, Y = 1)$, and define the analogous unlabeled target quantities $d_t = Q(S \ne R)$ and $\theta_t = Q(S = 1, R = 0 \mid S \ne R)$. Then $g_S - g_R = d_+(2\theta_+ - 1)$ and $\mu_S - \mu_R = d_t(2\theta_t - 1)$. Using six one-sided exact-binomial

bounds at $\delta/6$, define

$$\underline{\Delta}_g^{\mathrm{fac}} = \begin{cases} \underline{d}_+(2\underline{\theta}_+ - 1), & \underline{\theta}_+ > 1/2, \\ -1, & \text{otherwise}, \end{cases} \tag{28}$$

$$\overline{\Delta}_{\mu,+}^{\mathrm{fac}} = \max\{0, \overline{d}_t(2\overline{\theta}_t - 1)\}, \tag{29}$$

$$\Gamma_{\mathrm{fac}} = \underline{\mu}_R \underline{\Delta}_g^{\mathrm{fac}} - \overline{g}_R \overline{\Delta}_{\mu,+}^{\mathrm{fac}}. \tag{30}$$

Here the orientation intervals condition on the observed number of disagreements. With joint coverage at least $1-\delta$, $\Gamma_{\mathrm{fac}} > 0$ is sufficient for $F_{\beta,S} > F_{\beta,R}$ for every $\beta > 0$. This reparameterization avoids subtracting two separately widened rare-cell intervals; it uses the same samples, six risk allocations, and assumptions as Theorem 1.

Proof of Proposition 1. The source-positive disagreement count is binomial with parameter d_+. Conditional on observing $D_+ = d$ disagreements, the receiver-only count is binomial with d trials and parameter θ_+; hence its conditional Clopper–Pearson interval has coverage at least $1 - \delta/6$ for every d, and therefore also unconditionally. The same argument applies to (d_t, θ_t) on the target sample. On the joint six-bound event, if $\underline{\theta}_+ > 1/2$, monotonicity on the positive half-line gives

$$g_S - g_R = d_+(2\theta_+ - 1) \geq \underline{d}_+(2\underline{\theta}_+ - 1) = \underline{\Delta}_g^{\mathrm{fac}}.$$

If $\overline{\theta}_t < 1/2$, then $\mu_S - \mu_R \leq 0$ and zero bounds its positive part. Otherwise monotonicity gives $\max\{0, \mu_S - \mu_R\} \leq \max\{0, \overline{d}_t(2\overline{\theta}_t - 1)\}$. Substitution into the cross-multiplied difference from Theorem 1 proves the claim. A union bound over the six exact one-sided intervals gives joint coverage at least $1 - \delta$; no independence between the two factors is required.

Independent candidate selection and the conservative variant. Candidate choice need not consume a family-wise budget when it precedes certification. We use a disjoint labeled source selector to choose one receiver per class and then discard the unchosen candidates. Conditional on the selector, the retained receiver is fixed and independent of both the source certificate and target-rate sample, so Theorem 1 applies directly. This differs from examining several candidate certificates and deploying whichever passes, which does require family-wise correction.

The conservative implementation replaces $\underline{\Delta}_g$ by $\underline{g}_S - \overline{g}_R$ and uses five bounds at level $\delta/5$. It satisfies the same implication but may abstain even when the fully paired source-disagreement bounds contain enough evidence to certify a gain.

D.4 Rollout and bounded-drift extensions

Proposition D.6 (fresh-batch consequence with an explicit margin). Fix $\beta > 0$ and an external prevalence floor $q \geq q_{\min} > 0$. For an independent rollout batch of size N, let $L_\beta = 1 + \beta^2$ and $\eta_N(\alpha) = \sqrt{\log(10/\alpha)/(2N)}$. If

$$\beta^2 q_{\min} > L_\beta \eta_N(\alpha), \qquad \frac{q_{\min}\Gamma_{\mathrm{pair}}}{L_\beta} > \frac{4L_\beta\eta_N(\alpha)}{\beta^2 q_{\min} - L_\beta\eta_N(\alpha)}, \tag{31}$$

then a stricter gate that rolls out only when the displayed margin condition holds has probability at most $\delta + \alpha$ of deploying S with realized $\widehat{F}_{\beta,S} \leq \widehat{F}_{\beta,R}$ on that fresh batch. This extension uses the external floor and quantitative margin to convert population improvement into a fresh-batch guarantee.

Proof. Let $a_H = Q(Y = 1, H = 1)$ and $D_H = \beta^2 q + \mu_H$. On the confidence event used by Theorem 1, the cross-multiplied remainder is at least Γ_{pair}, so with $L_\beta = 1 + \beta^2$,

$$F_{\beta,S} - F_{\beta,R} \geq \frac{q_{\min}\Gamma_{\mathrm{pair}}}{L_\beta}, \tag{32}$$

because $D_R, D_S \leq L_\beta$. On a fresh batch, apply Hoeffding's inequality to the five empirical means for $q, a_R, a_S, \mu_R, \mu_S$. With probability at least $1 - \alpha$, all five errors are at most

$\eta_N(\alpha) = \sqrt{\log(10/\alpha)/(2N)}$. If $\beta^2 q_{\min} > L_\beta \eta_N(\alpha)$, elementary ratio perturbation gives, for each $H \in \{R, S\}$,

$$|\widehat{F}_{\beta,H} - F_{\beta,H}| \le \frac{2L_\beta \eta_N(\alpha)}{\beta^2 q_{\min} - L_\beta \eta_N(\alpha)}. \tag{33}$$

The displayed margin condition makes the certified population gap larger than the sum of these two deviations. A union bound over certificate failure and fresh-batch concentration gives failure probability at most $\delta + \alpha$.

Corollary D.7 (family-wise rollout). For a pre-declared family of M policy comparisons, assign risks $\delta_1, \ldots, \delta_M$ such that $\sum_{j=1}^{M} \delta_j \le \delta_{\mathrm{F}}$ and run comparison j with its allocated risk δ_j. With probability at least $1 - \delta_{\mathrm{F}}$, every approved comparison improves every F_β simultaneously. This Bonferroni construction needs no independence across classes or candidate models, but its wider intervals can increase abstention.

Proof. For comparison j, Theorem 1 implies that the probability of approving a non-improving policy is at most δ_j. A union bound over the pre-declared family gives probability at most $\sum_j \delta_j \le \delta_{\mathrm{F}}$ that any approved comparison is non-improving. No cross-comparison independence is required.

Proposition D.8 (bounded positive-conditional drift). Write $g_H^P = P(H = 1 \mid Y = 1)$ and $g_H^Q = Q(H = 1 \mid Y = 1)$ for $H \in \{R, S\}$. Replace exact positive-conditional stability by declared budgets $|g_H^Q - g_H^P| \le \epsilon_H$; the bounds on u_+, v_+ and g_R below are computed from source positives. Let

$$\underline{\Delta}_{g,\epsilon} = \underline{u}_+ - \overline{v}_+ - \epsilon_S - \epsilon_R, \tag{34}$$

$$\overline{g}_{R,\epsilon} = \min\{1, \overline{g}_R + \epsilon_R\}, \tag{35}$$

$$\Gamma_{\mathrm{drift}} = \underline{\mu}_R \underline{\Delta}_{g,\epsilon} - \overline{g}_{R,\epsilon} \max\{0, \overline{\Delta}_\mu\}. \tag{36}$$

On the same confidence event as Theorem 1, $\Gamma_{\mathrm{drift}} > 0$ is sufficient for $F_{\beta,S} > F_{\beta,R}$ on the target distribution for every $\beta > 0$. The budgets are externally justified policy-level robustness envelopes. Unlabeled action rates do not identify candidate-specific TPR drift. Increasing the declared budgets monotonically reduces approval coverage and can eliminate all positive margins.

Proof. The drift budgets imply $g_S^Q - g_R^Q \ge g_S^P - g_R^P - \epsilon_S - \epsilon_R \ge \underline{\Delta}_{g,\epsilon}$ and $g_R^Q \le \min\{1, \overline{g}_R + \epsilon_R\} = \overline{g}_{R,\epsilon}$. Target decision rates are observed on the unlabeled certificate sample exactly as in Theorem 1. Substituting these bounds into the cross-multiplied F_β difference and dropping the nonnegative $\beta^2 q(g_S^Q - g_R^Q)$ term proves the claim whenever $\Gamma_{\mathrm{drift}} > 0$.

Corollary D.9 (critical receiver-drift radius). Consider one-sided receiver TPR erosion: $\epsilon_R = 0$ and $g_S^Q \ge g_S^P - \rho$. Define the zero-drift certified margin

$$\Gamma_0 = \underline{\mu}_R(\underline{u}_+ - \overline{v}_+) - \overline{g}_R \max\{0, \overline{\Delta}_\mu\}. \tag{37}$$

If $\Gamma_0 > 0$ and $\underline{\mu}_R > 0$, let

$$\rho^\star = \frac{\Gamma_0}{\underline{\mu}_R}. \tag{38}$$

Then every declared budget $0 \le \rho < \rho^\star$ remains certified. The value $\rho^\star$ is an interpretable robustness margin—the amount of receiver TPR erosion that would exhaust the certificate—not an estimate of the unobserved target drift.

Proof. With $\epsilon_R = 0$ and $\epsilon_S = \rho$, Proposition D.8 gives $\Gamma_{\mathrm{drift}}(\rho) = \Gamma_0 - \underline{\mu}_R \rho$. This quantity is strictly positive exactly when $\rho < \Gamma_0/\underline{\mu}_R = \rho^\star$, proving the claim.

Specifying drift budgets. The assumption says that, among truly actionable cases, each frozen policy's capture probability transfers from the historical positive archive to current traffic. Stable label definitions and measurement pipelines, with changes primarily in prevalence or negative-case

composition, motivate this condition. Language, hospital, sensor, annotation, or attack-strategy changes can alter the positive population and require a drift allowance. Pre-declared budgets may draw on held-out historical TPR movements or expert-specified stress envelopes; their validity for the next target window remains an assumption. Delayed labels can inform budgets for subsequent windows. The natural-shift audits evaluate empirical behavior under unverified capture-rate stability.

D.5 Identifiability under conditional shift

Falsification versus validation. Under the stronger assumption that the entire four-cell distribution $(R, S) \mid Y$ is stable, source class-conditionals generate a one-dimensional mixture family indexed by target prevalence. A target joint-decision table outside this family can falsify that stronger joint label-shift model. Passing such a check cannot validate the positive-conditional stability used by C-RM: that weaker assumption fixes only the $Y = 1$ conditionals and leaves the target $Y = 0$ cells free. We therefore treat a four-cell compatibility check only as an optional diagnostic under the stronger model, not as part of the certificate.

Proposition D.10 (target-unlabeled identifiability boundary). Let R and S be fixed binary policies with equal target action rate $\mu_R = \mu_S$ and nonzero disagreement probability. Suppose a deployment gate observes arbitrary labeled source data and unlabeled target inputs, scores, and the joint decisions (R, S), but no target labels. For every $\beta > 0$, there exist two target distributions Q^+ and Q^- that induce exactly the same observable data and the same target prevalence, yet

$$F_{\beta,S}(Q^+) > F_{\beta,R}(Q^+), \qquad F_{\beta,S}(Q^-) < F_{\beta,R}(Q^-). \tag{39}$$

Any gate measurable only from these observables has the same approval probability a in the two worlds and harmful-approval probability a in Q^-. Thus a uniform erroneous-approval bound of δ requires $a \le \delta$ on this observational law. In particular, a zero-error guarantee requires abstention. Approving more often than δ requires additional information, such as a valid stability restriction, drift envelope, or target outcomes.

Even after RM removes score-scale and action-count differences, the sign of the local ranking gain requires additional information. C-RM links historical and target evidence through positive-class stability or a declared drift envelope.

Proof. Fix the target marginal Q_X and the two policies. Equal action rates imply that the S-only and R-only disagreement regions have equal mass; write $\pi = Q(S = 1, R = 0) = Q(S = 0, R = 1) > 0$. Define Q^+ by assigning $Y = 1$ exactly on the S-only region and $Y = 0$ elsewhere, and define Q^- by assigning $Y = 1$ exactly on the R-only region. Keep any labeled source distribution P fixed across the two worlds. Both worlds therefore have the same source law, target inputs, scores, joint decisions, action rate μ, and target prevalence $q = \pi$. Under Q^+, $TP_S = \pi$ and $TP_R = 0$; under Q^- the two values are reversed. Since $F_{\beta,H} = (1 + \beta^2)TP_H/(\beta^2 q + \mu_H)$ and the denominators agree, the asserted inequalities hold for every $\beta > 0$. Any gate based only on the observable data has the same output distribution in both worlds. Write a for its approval probability, including any internal randomization. Every approval is harmful under Q^-, so its erroneous-approval probability there is exactly a. A uniform bound of δ therefore requires $a \le \delta$; setting $\delta = 0$ proves the zero-error abstention claim.

E Additional Experimental Details

CivilComments architecture and training diagnostics. A three-seed diagnostic replaces the encoder with a source-fine-tuned Qwen3-0.6B decoder. It reaches mean fixed-count F1 $.6595 \pm .0013$; C-RM accepts 141/150 global partitions with no realized loss and retains the reference on low-support identity-group comparisons. This decoder replication supports the same evidence-adaptive behavior observed with encoder-only candidates.

RCV1 controlled shift. We use the standard LYRL order: the first 23,149 examples form training, and a fixed seed-0 permutation of the remaining pool supplies 60,000 validation examples and the target resampling pool. Classes are selected using training prevalence in $(0.002, 0.05)$ and more than 30 validation positives, then the 30 rarest eligible labels are retained. The reference is logistic

regression with $C = 1$; the receiving ranker averages percentile ranks from $C \in \{0.1, 1, 10\}$. Thresholds maximize validation F1 over 200 score quantiles from .5 to .9999. Each one-vs-rest target batch contains 20,000 examples, uses prior multipliers $\{1, 1.5, 2, 3, 5, 8\}$, and is repeated for 10 seeds. Platt calibration uses logistic regression on reference logits. This construction does not preserve cross-label co-occurrence.

GoEmotions deep transfer. We use the official simplified train/validation/test split and TF–IDF bigrams (minimum document frequency 2, sublinear term frequency). Class selection uses only training prevalence in $(0.001, 0.08)$ and at least 15 validation positives, yielding 20 non-neutral classes. RoBERTa-base is fine-tuned for three epochs with multilabel BCE and class-wise positive weights capped at 20. Target batches contain 5,000 examples and use 10 seeds. The fixed-count comparison evaluates both rankers at exactly the reference-derived k. For the strict model-upgrade audit in Table 3, a fixed seed-20260820 permutation divides the 5,426 official validation examples into 1,085 selector/checkpoint and 4,341 source-certificate examples. Support requirements leave 25 classes. We use prior multipliers $\{.5, 1, 1.5, 2, 3, 5, 8\}$ and 50 target draws of size 5,000 per receiver seed. Reference thresholds and receiver checkpoints are frozen on the selector; the gate uses source-certificate labels and unlabeled target decisions only. For each class, shift, and seed, the implementation draws separate target matching, rate-certification, and evaluation samples. The candidate cutoff is set on the matching sample and then applied unchanged to certification and evaluation. Population F1 is computed under the empirical class-conditional resampling distribution; finite-batch F1 uses the separate evaluation draw. Deterministic sub-gap tie breaking enforces exact Top-k cardinality even when target resampling repeats an example in the matching window. The 20-class deep-transfer panel uses the global prevalence/support filter; the 25-class audit uses the frozen selector/certificate split and its corresponding support criterion.

CIFAR-100 deep transfer. We split the official training images into 80% probe training and 20% validation, preserving the official test set as the target resampling pool. ImageNet-pretrained encoders are frozen and three pairs of independently initialized linear 100-way heads are trained for 30 epochs. Reference validation thresholds are selected class by class. Each class is then evaluated on ten 5,000-example target resamples at prior multipliers $\{0.5, 1, 2, 3, 5, 8\}$.

COCO deep transfer. We use all 36 COCO-format parquet shards, giving 117,266 images, and a fixed 70/5/5/20 split for head training, threshold calibration, source-local selection, and target resampling. Frozen ImageNet ResNet-50 and DINOv2-base features receive independent 80-way linear heads. Results average three head seeds, 30 source-selected labels, ten target seeds, and 5,000-image bags. The primary protocol resamples each class independently. A second protocol enriches the label whose source prevalence is closest to .08 and evaluates all 30 labels on the same sampled images, preserving their observed co-occurrence.

Certified-transfer protocol. The primary population-certificate study uses all 117,266 labeled images and a fixed 50/5/25/20 split for head training, reference-threshold selection, source-positive certification, and the target resampling pool. Its 30-class panel was fixed by the preceding ten-shard study before the full-data result. Each controlled target draw is itself separated into threshold, rate-certificate, and evaluation samples. Population F1 is computed exactly under the finite empirical class-conditional resampling distribution; finite-batch F1 is a separate diagnostic. Confidence, source-support, target-size, corruption, and five-split ablations use the 32,579-image archive with the same 30 classes. With the 25% source certificate fixed, target subset sizes 500/1,000/2,000/5,000 produce 7/19/49/166 switches on the full archive and the same 100 target seeds, with no population loss and finite-batch losses of 2/0/0/0, confirming that certificate power improves as unlabeled evidence accumulates.

The switch count is identical because both paired gates certify the beta-independent positive remainder obtained after dropping $\beta^2 q(g_S - g_R)$ in Theorem 1. Recomputing the deployed objective confirms that every approved comparison is non-losing for precision-emphasizing, balanced, and recall-emphasizing choices of β.

Strong quantification controls. For GoEmotions, all count estimators share the rank average of the same three RoBERTa runs. PACC and MLLS use Platt-calibrated TF–IDF reference scores. RLLS regularization is selected per class using only simulated enrichments of the source validation

Table 9: F_β sensitivity of factorized C-RM on full COCO. Gains compare receiver and reference at the same transported count; every reported switch is non-losing on the population objective.

β	Fixed-count gain, $1\times$	Fixed-count gain, $8\times$	Switches/losses
.5	+.0725	+.0184	1,334/0
1	+.0711	+.0169	1,334/0
2	+.0706	+.0156	1,334/0

split. DFM matches 128-dimensional random Fourier feature means of calibrated reference scores. Target labels are used only to construct controlled resamples and evaluate F1; no target label selects a method or hyperparameter.

The locally auditable evidence register, raw arrays, split logic, and generated CSV tables accompany the code.