

The Calibration Transfer Gap: When Recalibration Parameters Fail to Transfer under Distribution Shift

Bheema Shanker Neyigapula

bheemashankerneyigapula@gmail.com

Independent Researcher

Research Article

Keywords: calibration, distribution shift, selective prediction, temperature scaling, trustworthy machine learning

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11014477/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

The Calibration Transfer Gap: When Recalibration Parameters Fail to Transfer under Distribution Shift

Bheema Shanker Neyigapula^{1*}

^{1*}Independent Researcher, Hyderabad, India.

Corresponding author(s). E-mail(s): bheemashankerneyigapula@gmail.com;

Abstract

Predictive uncertainty is known to degrade under distribution shift. We ask a sharper question: when calibration breaks, is it the recalibration *method* that fails, or its fitted *parameter* that fails to transfer? We isolate the latter with the *calibration transfer gap*, the excess calibration error from transferring a source-fitted calibrator versus refitting on the target, measured with disjoint target-calibration and target-test splits. Across models from linear classifiers to fine-tuned transformers and three calibrator families (temperature, Platt, isotonic), the gap is *regime-dependent*. Under sentiment-domain shift it is negligible in-domain but widens significantly (temperature expected-calibration-error gap **0.003** $\rightarrow$ **0.021**), robustly across binning schemes and scoring rules; a competent cross-genre inference model shows no detectable gap, and a tabular benchmark with natural geographic shift reproduces both regimes. Recalibration-parameter drift (the log-ratio of fitted temperatures) is associated with the gap in shared-fit comparisons (text $\rho = \mathbf{0.75}$, tabular $\rho = \mathbf{0.63}$), but the association weakens under independent estimation ($\rho = \mathbf{0.64} \rightarrow \mathbf{0.35}$) and varies across seeds. We prove an identifiability limit: mean confidence, variance, and accuracy do not determine the conditional-correctness curve, and several label-free strategies indeed fail to beat the source calibrator. Finally, for the decision of whether to refit given a small target-label budget, unconditional refitting on a few dozen labels is often harmful; uncertainty-aware rules (shrinkage, cross-validation, an information-criterion gate) reduce but do not eliminate that harm and at the smallest budgets still fail to beat keeping the source. In-domain calibration is thus not a sufficient deployment guarantee under substantial confidence shift. Experiments are reproducible from supplementary scripts.

Keywords: calibration, distribution shift, selective prediction, temperature scaling, trustworthy machine learning

1 Introduction

Machine-learning systems are increasingly trusted not merely to predict, but to report *how sure* they are. A model’s confidence drives selective prediction and abstention, human-in-the-loop escalation, cost-efficient routing between models, and risk-sensitive decisions in medicine and finance. For these uses to be safe, confidence must be

calibrated: among inputs a model labels with probability p , a fraction close to p should be correct (Guo et al. 2017). A large literature accordingly studies calibration and post-hoc fixes such as temperature scaling (Guo et al. 2017; Platt 1999).

Almost all of this evidence is gathered *in-domain*: calibration is measured on held-out data from the training distribution. Deployment is rarely so tidy. A sentiment model trained on

movie reviews meets product reviews; an inference model trained on one text genre meets another. Under such *distribution shift*, accuracy is known to fall (Quiñero-Candela et al. 2009; Koh et al. 2021), and Ovadia et al. (2019) document that predictive uncertainty deteriorates for deep models. That calibration degrades under shift is, by now, expected. The scientifically sharper questions, and the ones a practitioner using post-hoc recalibration actually faces, are *why* it degrades and, crucially, *when*: is degradation universal, or does it depend on the task and model? We will show it is regime-dependent, so identifying the regime becomes the central problem.

Method failure versus parameter-transfer failure.

Post-hoc recalibration has two parts: a functional *family* (e.g. dividing logits by a scalar temperature) and a *parameter* within it (the value T) fitted on held-out source data. When calibration breaks under shift, either the family is inadequate on the target, or the family is fine but the *source-fitted parameter* is wrong for the target. These have opposite implications: the first says we need richer recalibrators, the second says we need to *re-estimate parameters on the target*, and that the barrier is doing so without target labels. Source-fitted versus target-fitted calibration comparisons already exist in the domain-adaptation literature: Pampari and Ermon (2020) compare a source-fitted isotonic calibrator against target-label calibration under covariate shift, using separate target calibration and evaluation subsets and studying shift magnitude and calibration sample size; Wang et al. (2020) include temperature scaling with a target-label oracle; and Gong et al. (2021) evaluate source-only versus target-only oracle calibration. We do not claim to introduce that comparison, the separate evaluation split, or the sample-size experiment. Our contribution is a consistent, leakage-controlled *characterization* across a fixed grid of tasks, models, and three calibrator families: we show the gap is regime-dependent and ordered by a single quantity (recalibration-parameter drift), quantify how much a same-sample target fit exaggerates it (a measurement audit), map the limited portability of unlabeled

warning signals, and pair a label-free identifiability limit with an independently-evaluated target-label budget. As a measurement tool we use the **calibration transfer gap**, with T_{src} the temperature fitted on a source-calibration split and T_{tgt} fitted on a *disjoint* target-calibration split within the same scalar family under the same objective (both by NLL; Section 3),

$$\Delta_{\text{transfer}} = \text{ECE}(T_{\text{src}}) - \text{ECE}(T_{\text{tgt}}), \quad (1)$$

the excess calibration error associated with transferring a source-fitted recalibration parameter relative to refitting within the same family on the target. Because both temperatures are fitted for NLL rather than for ECE, Δ_{transfer} is a *signed* descriptive difference, not a nonnegative decomposition of calibration error: it can in principle be negative if the target-NLL temperature happens to raise ECE, and we report it as such (in-domain values near zero are sometimes slightly negative). A large positive Δ_{transfer} identifies a component of the calibration error associated with parameter transfer while holding the recalibration family fixed, which reframes the question from “is confidence unreliable under shift?” to “*when* must the recalibration parameter be re-estimated on the target, and can mean-level label-free signals do it?”

Contributions.

We make three contributions.

- **Calibration transfer is regime-dependent.**

We define the transfer gap (Eq. 1) as a measurement tool and use it to map *when* calibration fails to transfer, via a *train-on-source, test-on-target* protocol with four strictly disjoint data roles, source-train, source-calibration, target-calibration, and an untouched target-test set, so that the target-fitted reference is never fit and evaluated on the same examples (Section 3). On sentiment-domain shift the gap is negligible in-domain but widens significantly under shift for every model family, from linear classifiers to a fully fine-tuned transformer, and for three calibrator families of increasing flexibility (temperature, Platt, isotonic), establishing that parameter transfer is a substantial source of calibration failure, not merely a limitation of scalar temperature scaling, robustly across

ECE bin counts, adaptive ECE, Brier, and NLL. On cross-genre NLI with a competent fine-tuned model (RoBERTa) target refitting yields no detectable average improvement, and on an independent tabular benchmark with natural geographic shift (Folktables ACSIncome) a single benchmark spans both regimes, delimiting the phenomenon as regime-dependent rather than universal and providing evidence against a text-specific artifact (Sections 3–6).

- **Characterization and downstream consequences.** We identify what predicts the gap: recalibration-parameter drift $|T_{\text{src}} - T_{\text{tgt}}|$, measured scale-invariantly as $|\log(T_{\text{tgt}}/T_{\text{src}})|$, is the strongest descriptive correlate in the shared-fit comparisons, associated with the gap at $\rho = 0.75$ on text and $\rho = 0.63$ on tabular data; confidence-distribution divergence is a usable but regime-dependent label-free proxy. A controlled shift-severity experiment examines the associations among shift severity, confidence drift, parameter drift, and the gap within one setting, and an independent-estimation analysis (fitting the drift and the gap from disjoint calibration halves, with a no-shift sampling reference) shows that the drift-gap association weakens at 125 examples per fit and varies across seeds, indicating sensitivity to calibration estimation. We connect the gap to selective prediction with a leakage-free threshold-transfer test, where an in-domain-tuned coverage policy admits roughly twice its intended error rate under shift (Sections 4–5).
- **Limits, and cost, of label-free recovery.** We prove that mean confidence, variance, and accuracy do not identify the conditional-correctness curve (Propositions 1–2), so no rule based on those statistics can in general recover a calibrated map that preserves useful confidence variation. Consistent with this, three label-free heuristics fail to beat the source calibrator on both linear models and a fine-tuned transformer, and two importance-weighted baselines, a confidence-only one on the transformer and a stronger feature-based one on the tabular benchmark (cross-fitted density ratios on the original features), likewise do not. As an actionable counterpart, a target-label-budget sweep shows that 50–100 labelled target examples recover roughly half (52–57%) of the average

improvement achieved by a 250-label reference in the linear sentiment setting, and a keep-or-refit analysis shows that unconditional refitting on a few dozen labels is often harmful; uncertainty-aware rules (shrinkage, cross-validation, or an information-criterion rule specialized to temperature refitting) reduce but do not eliminate that harm, and at the smallest budgets still fail to beat keeping the source calibrator, a meaningful negative result rather than a solved decision (Sections 7–7.4).

All datasets are public and all experiments are accompanied by scripts and fixed seeds provided as supplementary material for reproduction; the definition, diagnostic, and identifiability results are model-agnostic.

2 Related Work

Calibration and its measurement.

Modern networks are frequently miscalibrated and typically overconfident, and temperature scaling substantially reduces in-domain Expected Calibration Error (Guo et al. 2017). Classical post-hoc alternatives include Platt scaling (Platt 1999) and isotonic regression (Zadrozny and Elkan 2002), while multiclass extensions such as vector, matrix, and Dirichlet calibration add parameters (Guo et al. 2017; Kull et al. 2019). Measurement is itself contested: binned ECE is biased and sensitive to bin count and binning scheme (Nixon et al. 2019; Roelofs et al. 2022), adaptive (equal-mass) binning (Nguyen and O’Connor 2015) and debiased or kernel estimators (Kumar et al. 2019; Zhang et al. 2020) reduce these artifacts, and proper scoring rules (Brier, NLL) provide binning-free summaries (Brier 1950; DeGroot and Fienberg 1983). We therefore report the transfer gap across several bin counts, adaptive ECE, Brier, and NLL rather than relying on a single 15-bin ECE. All these methods are ordinarily fit and evaluated in-distribution; we ask what happens to such a fit under shift.

Calibration under distribution shift.

Robustness to shift is long-standing (Quiñonero-Candela et al. 2009), with benchmarks documenting large out-of-distribution accuracy drops (Koh et al. 2021; Hendrycks and Dietterich 2019). Most directly, Ovadia et al. (2019) show that predictive uncertainty, including post-hoc-calibrated

uncertainty, degrades under shift for deep models, and [Minderer et al. \(2021\)](#) revisit calibration across modern architectures. A line of work seeks to *maintain* calibration under shift using unlabeled target data, via importance-weighted recalibration under covariate shift ([Park et al. 2020](#)), calibration domain generalization ([Gong et al. 2021](#)), calibration under prior/label shift ([Alexandari et al. 2020](#); [Lipton et al. 2018](#)), and transferable calibration for domain adaptation ([Wang et al. 2020](#)). Several of these already compare source-fitted against target-fitted (or target-oracle) calibration. Most directly, [Pampari and Ermon \(2020\)](#) compare a source-fitted isotonic calibrator against target-label calibration under covariate shift, and their study already includes separate target calibration and evaluation subsets and analyses of shift magnitude and calibration sample size; [Wang et al. \(2020\)](#) use temperature scaling with a target-label oracle, and [Gong et al. \(2021\)](#) evaluate source-only versus target-only oracle calibration. Our aim is therefore not to introduce the source-versus-target comparison, the disjoint evaluation split, or the sample-size analysis, all of which have precedents, but to characterize the phenomenon systematically. Relative to these works we (i) run one consistent, leakage-controlled comparison across a fixed task/model/calibrator grid rather than as a reference baseline for a proposed method, (ii) show the comparison is regime-dependent and that a single quantity, recalibration-parameter drift, orders the regimes across text and tabular data, (iii) quantify how much a same-sample target fit exaggerates the apparent benefit of refitting, and (iv) pair a label-free identifiability limit with an independently-evaluated target-label budget. We verify the pattern persists across temperature, Platt, and isotonic families.

Label-free performance and risk estimation.

Estimating target accuracy without labels is possible from confidence statistics: Average Thresholded Confidence (ATC) ([Garg et al. 2022](#)), confidence- and distribution-based estimators ([Guillory et al. 2021](#)), and agreement-based estimators ([Jiang et al. 2022](#)). Label shift can be corrected from unlabeled data via BBSE and related estimators ([Lipton et al. 2018](#); [Alexandari](#)

[et al. 2020](#)). We use ATC to drive label-free recalibration and show, both empirically and via an identifiability argument, that an accurate accuracy estimate does not by itself determine calibration, explaining why such signals are insufficient for recalibration under shift.

Uncertainty-aware and label-efficient calibration.

Several lines of work use uncertainty or limited labels for calibration. [Clarté et al. \(2023\)](#) match average confidence to accuracy (expectation consistency) as a calibration objective with theoretical backing; [Ji et al. \(2021\)](#) assess a classifier’s calibration label-efficiently via active Bayesian sampling; and [Angelopoulos et al. \(2021\)](#) select among prediction procedures with finite-sample risk control. Our keep-or-refit analysis (Section 7.4) is narrower and distinct: it targets the specific binary decision of whether a small target-label budget should be spent re-estimating a *source-fitted* calibrator under shift, quantifies the finite-label estimation cost through a bias-variance approximation, and evaluates the decision on held-out domains. We do not claim to introduce uncertainty-aware calibration; we characterize when limited-label refitting of a transferred calibrator is worthwhile, and find a derived criterion and simpler shrinkage/cross-validation baselines perform comparably.

Selective prediction and abstention.

Selective classifiers abstain when uncertain ([El-Yaniv and Wiener 2010](#); [Geifman and El-Yaniv 2017](#)), trading coverage for risk along the risk-coverage curve ([Geifman and El-Yaniv 2019](#)), whose area (AURC) summarizes selective-prediction performance across coverage levels and depends on both confidence ordering and the overall error rate. Robustness of selective prediction and OOD detection under shift is an active concern ([Kamath et al. 2020](#); [Hendrycks and Gimpel 2017](#)). We use risk-coverage, and a source-to-target threshold-transfer test, to measure whether confidence remains *useful* for abstention under shift, beyond numeric calibration.

3 Method

Setup and protocol.

We study two classification tasks whose domains share a prediction task and label space, so cross-domain evaluation is a natural domain shift rather than a change of task. We do not claim pure covariate shift (i.e. that $P(Y | X)$ is preserved), which shared labels alone do not establish; the shifts here may include changes in the label-conditional distribution. **Sentiment** (binary) uses four domains: SST-2, IMDB, Yelp Polarity, and Rotten Tomatoes. **Natural-language inference** (three classes: entailment, neutral, contradiction) uses five MNLI genres: slate, telephone, government, fiction, travel. We use four strictly disjoint data roles to avoid any in-sample optimism in the target-fitted reference: (i) a *source-training* split fits the predictive model; (ii) a held-out *source-calibration* split fits the transferred (source) calibrator; (iii) a separate *target-calibration* split fits the target-fitted reference calibrator; and (iv) an *untouched target-test* split on which both calibrators are evaluated. The source- and target-calibration splits are matched in size within each setting ($N_{\text{cal}} = 250$ for the linear, tabular, and fine-tuned-DistilBERT experiments; $N_{\text{cal}} = 400$ for the three-class RoBERTa MNLI experiment, where more examples are needed to fit a three-class temperature stably) so that the source-versus-target comparison isolates the effect of *shift* rather than of calibration sample size; we separately study the effect of the number of target-calibration labels in the budget sweep of Section 7.3. Exact per-setting split sizes are in Table A1. Evaluating source and target calibrators on the same untouched test split makes the transfer gap a like-for-like comparison free of the leakage that arises when a calibrator is fit and scored on the same examples (isotonic regression being especially vulnerable). One experiment is an explicit exception to label-free adaptation: the ACS importance-weighting baseline (Section 7) uses the target-test *features* for transductive domain-classifier training while their *labels* remain reserved for evaluation; every other experiment uses no target-test information of any kind for fitting. Each source-target combination forms a cell of a shift matrix whose diagonal is in-domain. The number of random seeds varies

by setting (five for the linear and tabular experiments, three for the fine-tuned transformers), as specified in the experimental details, and we report 95% confidence intervals. Exact dataset sizes, hyperparameters, and fitting procedures are given in Appendix A.

Model spectrum.

To test whether findings are artifacts of weak models, we span four model families: TF-IDF with ℓ_2 logistic regression (LR); TF-IDF with a calibrated linear SVM (LSVM); frozen DistilBERT (Sanh et al. 2019) [CLS] embeddings with a logistic-regression head trained per source; and *end-to-end fine-tuned* DistilBERT, all layers updated on each source domain. For the NLI task we additionally fine-tune *RoBERTa-base* (Liu et al. 2019) end-to-end per source genre, giving a competent modern classifier well above the three-class floor, so that the NLI evidence does not rest on a near-chance model. We note that the LSVM already applies Platt-style probability calibration internally (its base scores are not probabilistic); our source/target recalibration is therefore measured *on top of* that in-training calibration for this model. This is a realistic deployment scenario (many probabilistic classifiers are already internally calibrated), but it means the LSVM measures transfer of a post-hoc layer applied to an already-calibrated model. The LR, frozen-feature, and fine-tuned models have no such internal calibration, and the transfer gap appears for all of them, so the effect is not an artifact of double calibration.

Confidence, calibration, and the transfer gap.

Let $P(x)$ be the predicted class-probability vector, $\hat{y} = \arg \max_k P_k$, and confidence $s = \max_k P_k$. We measure calibration with Expected Calibration Error (ECE) using 15 equal-width bins,

$$\text{ECE} = \sum_{b=1}^B \frac{|B_b|}{n} |\text{acc}(B_b) - \text{conf}(B_b)|. \quad (2)$$

Temperature scaling divides logits by a scalar $T > 0$ before the softmax, with T minimizing negative log-likelihood (NLL) (Guo et al. 2017). We fit $T_{\text{src}}^{\text{NLL}}$ on the source calibration split and apply it to every target, as in realistic deployment. To

separate the limits of the recalibration *family* from the limits of *parameter transfer*, we also fit a target temperature $T_{\text{tgt}}^{\text{NLL}}$ by minimizing the same NLL objective on target labels. The *calibration transfer gap*

$$\Delta_{\text{transfer}} = \text{ECE}(T_{\text{src}}^{\text{NLL}}) - \text{ECE}(T_{\text{tgt}}^{\text{NLL}}) \quad (3)$$

is the excess calibration error associated with source-to-target parameter transfer, relative to refitting the temperature on target labels within the same family under the same objective. We stress the scope of $T_{\text{tgt}}^{\text{NLL}}$: it is the temperature fitted on target labels by minimizing NLL, the standard temperature-scaling objective. It is *not* the ECE-minimizing scalar (which would require target labels and a different, non-standard objective), so it is a realistic within-family target reference rather than a general lower bound on achievable target ECE. A richer family (e.g. vector or matrix scaling), or direct ECE minimization, could in principle attain lower target ECE; a large Δ_{transfer} therefore establishes that standard scalar recalibration *can* substantially improve calibration on the target and that source-to-target parameter transfer contributes substantially to the observed calibration degradation, not that no richer method could beat $T_{\text{tgt}}^{\text{NLL}}$.

Calibrator families.

To check that the transfer gap is not specific to scalar temperature, we repeat the source-fit-versus-target-fit comparison for three calibrator families of increasing flexibility: temperature scaling (one parameter), Platt scaling (an affine map $az + b$ on the logit z , two parameters), and isotonic regression (a non-parametric monotone map). Each is fit on the source calibration split and, as a within-family reference, on target labels, and Δ_{transfer} is computed per family. This multi-calibrator comparison is conducted on the *binary sentiment* task, where all three families have an unambiguous one-dimensional form; the affine and isotonic maps act on the positive-class probability. Extending Platt and isotonic scaling to the multiclass NLI setting (e.g. via vector scaling or one-vs-rest isotonic) raises separate design choices and is out of scope here, so for NLI we report the temperature family only. A subtlety distinguishes these families from temperature scaling: temperature is a strictly monotone transform of

the logit and so never changes the predicted label or accuracy, whereas Platt and isotonic maps can move the positive-class probability across the 0.5 decision boundary and thus flip predictions. We measured this: across the sentiment shifted pairs the source-Platt and source-isotonic maps change the argmax label for on average 2.5% and 5.9% of target examples respectively (though the mean accuracy *change* is small and slightly positive, +0.002 and +0.003). The Platt and isotonic transfer gaps therefore reflect a mixture of calibration change and a modest amount of altered classification, whereas the temperature-family gap, our primary quantity, is a pure calibration effect (temperature never flips a binary prediction). We report per-pair flip rates in the supplementary logs and rely on the temperature family for the headline claims for exactly this reason.

Selective prediction.

We rank test inputs by confidence and, at coverage c , retain the most confident fraction c . The risk-coverage curve plots selective error against c ; we report its area (AURC, lower is better). AURC summarizes selective-prediction performance across coverage levels and depends on both the confidence ordering and the overall error rate.

Statistical treatment.

The experimental unit is nested: task $\rightarrow$ source $\rightarrow$ target $\rightarrow$ model $\rightarrow$ seed, and cells sharing a seed and model are not independent. We adopt a single, consistently applied uncertainty procedure: all effect sizes (e.g. the change in mean transfer gap) are reported with 95% confidence intervals from a *seed-level bootstrap*, i.e. we first aggregate each quantity to one value per seed (averaging over the dependent source-target cells and models within that seed) and then bootstrap over the independent seeds. This respects the dependence structure by never resampling correlated cells independently; the per-seed aggregation absorbs the within-seed grouping that a naive cell-level bootstrap would violate. Spearman correlations are computed over the pooled shifted pairs and reported as *descriptive* summaries only, accompanied by the range of the coefficient recomputed within each seed (a dependency-preserving check); we do not attach pooled-pair bootstrap confidence

intervals to them, since the pooled pairs share source models and target sets. We prefer effect sizes with CIs to raw p -values throughout, and we stress that the seed-level intervals describe variation across repeated runs on the fixed datasets and domains studied, not precision about performance on arbitrary new domains.

Label-free recalibration (three strategies).

Since the target is unlabeled in practice, we ask whether the transfer gap can be closed without target labels, testing three strategies. (i) *Mean-matching*: estimate target accuracy label-free with ATC (Garg et al. 2022), then choose T so target mean confidence equals this estimate. (ii) *Pseudo-ECE*: assign each target point a pseudo-correctness label $\tilde{c}(x) = \mathbb{I}[s(x) > t]$ from the ATC threshold t , and minimize a distribution-level pseudo-ECE against $\tilde{c}$. (iii) *Variance-matching*: choose T so target confidence variance matches the source. We test whether any reliably improves ECE over the source temperature across seeds and tasks.

4 Results

Accuracy falls under shift.

On sentiment, accuracy is 0.788 ± 0.004 in-domain and 0.714 ± 0.004 under shift (a 7.4-point drop) for the linear models, and $0.839 \pm 0.004 \rightarrow 0.799 \pm 0.006$ for fine-tuned DistilBERT. For cross-genre NLI, fine-tuned RoBERTa reaches mean in-domain accuracy 0.659 (95% CI [0.605, 0.694]) and 0.634 under shift ([0.577, 0.671]; Table 2), well above the three-class floor.

Existence on sentiment: the recalibration parameter fails to transfer.

This is our existence result on sentiment (Table 1). Temperature scaling behaves as expected in-domain, reducing ECE (from 0.049 raw to 0.035) with a negligible transfer gap of 0.003 (95% CI $[-0.001, 0.006]$, i.e. statistically indistinguishable from zero); under shift the same source temperature leaves ECE at 0.054, no better than uncalibrated. The decisive comparison is that the target-fitted temperature, refit on the disjoint target-calibration split and scored on the untouched target-test set, reaches 0.032: within the same scalar family a good parameter still

exists on the target, showing that parameter transfer accounts for a substantial component of the shifted calibration error. The transfer gap widens from 0.003 in-domain to 0.021 under shift (95% CI [0.015, 0.029]), a paired increase of $+0.019$ [0.009, 0.030] with a CI excluding zero (Figure 2). The underlying miscalibration is systematic across domain pairs: the uncalibrated ECE is on average higher for off-diagonal (shifted) pairs (0.053) than for the in-domain diagonal (0.049), and the raw-ECE matrix (Figure 3) shows the shifted cells are typically, though not uniformly, worse.

On sentiment the effect is *not* an artifact of weak models, though it attenuates as the model strengthens. With frozen DistilBERT features the shifted transfer gap is 0.045, an order of magnitude above its in-domain value (≈ 0.002). With *fully fine-tuned* DistilBERT, reaching 0.84 in-domain accuracy, temperature scaling is near-perfect in-domain (gap ≈ 0) yet leaves a small but significant residual under shift (gap 0.007, 95% CI [0.005, 0.009]). The phenomenon thus holds, in diminishing magnitude, across the encoder-classifier spectrum from linear models to a fine-tuned transformer on this task family (Table 3); we do not claim it for larger pretrained or generative models.

The gap is regime-dependent: cross-genre NLI with a strong model.

Whether the gap generalizes to a different task and a genuinely capable classifier is a separate question, and the answer sharpens the contribution. We fine-tune RoBERTa-base end-to-end per MNLI genre (in-domain accuracy 0.60–0.69 across seeds, well above the 0.33 three-class floor and the near-chance linear model this replaces) and measure the cross-genre transfer gap over three seeds under the same leakage-free protocol, fitting the target temperature on a disjoint target-calibration split and scoring on the untouched target-test set (Table 2). The deployable quantity, the *source-fitted* temperature’s ECE, is essentially unchanged in-domain versus shifted ($0.057 \rightarrow 0.064$), and, decisively, the target-fitted temperature is *no better* under shift (0.065): the shifted transfer gap is -0.000 (CI $[-0.005, 0.003]$), indistinguishable from zero, versus 0.021 on sentiment. The paired shifted-minus-in-domain change is $+0.002$

(CI $[-0.002, 0.005]$), containing zero. For this competent NLI model there is therefore no detectable average improvement from target refitting under this protocol: refitting the temperature on target labels does not, on average, lower ECE, so the source parameter transfers adequately here.

This is an important boundary case. We state it carefully: under this protocol the target-fitted reference shows *no detectable average improvement* over the source calibrator (shifted gap -0.000 , CI $[-0.005, 0.003]$), so we do not detect a transfer gap rather than proving its exact absence. We do not claim formal statistical equivalence, which would require a pre-specified margin we did not set; we report only that the estimated average gap and its interval are small (and far below the sentiment effect) at this sample size. Consistent with the characterization in Section 4, the optimal temperature moves less across MNLi genres than across sentiment domains (Table 7). The transfer gap is thus not a universal law of distribution shift but a regime-dependent phenomenon. We use a competent fine-tuned model here rather than a near-chance linear NLI classifier, whose confidence would be too weak to make the calibration comparison meaningful.

The gap is not specific to temperature scaling.

A natural objection is that the transfer gap merely reflects the rigidity of a single scalar temperature. It does not. We repeat the source-fit-versus-target-fit comparison for three calibrator families of increasing flexibility (Table 4): temperature $0.003 \rightarrow 0.021$, Platt $0.001 \rightarrow 0.020$, isotonic $0.006 \rightarrow 0.035$, with paired shifted-minus-in-domain increases of $+0.019$ (95% bootstrap CI $[0.009, 0.030]$), $+0.019$ $[0.013, 0.026]$, and $+0.030$ $[0.024, 0.036]$ respectively. More flexible calibrators do not rescue transfer, and isotonic regression, the most flexible, has the largest gap, consistent with its greater sensitivity to the calibration set. We note that isotonic’s gap in particular shrinks sharply relative to a naive in-sample estimate: under the leakage-free split its shifted gap is 0.035 rather than the 0.070 obtained when the target isotonic map is fit and scored on the same examples, underscoring why the disjoint-split protocol matters most for the most flexible family. These results show that parameter transfer constitutes a substantial source of calibration failure under

shift and cannot be explained solely by the limited flexibility of scalar temperature scaling. Robustness across calibration metrics is addressed in Section 4.

Robustness across calibration metrics.

The central finding does not depend on 15-bin ECE. Table 5 reports the source-fit-minus-target-fit difference under each metric, in-domain and under shift, with the paired increase and its 95% seed-level bootstrap CI, across equal-width ECE at several bin counts, adaptive (equal-mass) ECE, Brier score, and NLL. We note that the target-fitted temperature is optimized for NLL, so for Brier and NLL these are source-fit-minus-target-fit metric *differences* rather than an oracle gap specific to each metric; they nonetheless provide binning-free robustness evidence. The shifted increase is positive with a CI excluding zero for every metric. Table 5 lists equal-width ECE at 10, 15, and 20 bins alongside adaptive ECE, Brier, and NLL; the two additional bin counts (30 and 50) behave identically (shifted-minus-in-domain increase $+0.021$, CI $[0.014, 0.029]$, and $+0.020$, CI $[0.013, 0.028]$) and are reported in the supplementary results file.

Consequences: selective prediction degrades.

Even imperfectly calibrated confidence may still support abstention if it ranks predictions well. Selective-prediction performance, measured by AURC, also deteriorates under shift: on sentiment, mean AURC rises from 0.100 in-domain to 0.162 shifted (Figure 4). We are explicit that this is a property of the model’s *raw* confidence, not of recalibration transfer: for binary classification a positive temperature T maps confidence to $\sigma(|z|/T)$, a strictly monotone transform of $|z|$, so it preserves the confidence ordering and leaves AURC on a fixed target set unchanged (up to ties). Refitting a temperature therefore cannot change AURC, so this deterioration is a separate, complementary failure to the calibration transfer gap. We note that AURC also depends on the overall error rate, so its increase reflects both any change in ranking quality and the higher shifted error rate; we therefore claim only that selective-prediction performance degrades, not that ranking quality alone deteriorated. The relevant deployment test

Table 1 Sentiment aggregate results with 95% confidence intervals over five seeds (linear models). ECE is reported uncalibrated (raw), with the source-NLL-fitted temperature ($T_{\text{src}}^{\text{NLL}}$), and with the target-NLL-fitted temperature ($T_{\text{tgt}}^{\text{NLL}}$); AURC is the risk–coverage area.

Metric (mean $\pm$ 95% CI)	In-domain	Shifted
Accuracy	0.788 ± 0.004	0.714 ± 0.004
ECE (raw)	0.049 ± 0.004	0.053 ± 0.003
ECE ($T_{\text{src}}^{\text{NLL}}$)	0.035 ± 0.003	0.054 ± 0.006
ECE ($T_{\text{tgt}}^{\text{NLL}}$)	0.032 ± 0.006	0.032 ± 0.002
Transfer gap	$0.003 [-0.001, 0.006]$	$0.021 [0.015, 0.029]$
AURC	0.100	0.162

Table 2 Cross-genre MNLI with fine-tuned RoBERTa-base, under the leakage-free protocol (target temperature fit on a disjoint target-calibration split, scored on the untouched target-test set; three seeds, 95% bootstrap CIs over seeds). Unlike sentiment, the source-fitted temperature’s ECE is essentially unchanged under shift and the target-fitted temperature is no better, so the transfer gap is indistinguishable from zero, delimiting the phenomenon as regime-dependent.

Metric (mean [95% CI])	In-domain	Shifted
Accuracy	0.659 [0.605, 0.694]	0.634 [0.577, 0.671]
ECE ($T_{\text{src}}^{\text{NLL}}$)	0.057 [0.050, 0.067]	0.064 [0.051, 0.073]
ECE ($T_{\text{tgt}}^{\text{NLL}}$)	0.060 [0.051, 0.076]	0.065 [0.051, 0.077]
Transfer gap	$-0.003 [-0.009, 0.002]$	$-0.000 [-0.005, 0.003]$

is thus the threshold-transfer experiment below, which concerns the confidence *level* at which one abstains, not the ranking.

To make the deployment consequence concrete, we run a leakage-free *threshold-transfer* test that mirrors practice: choose a confidence threshold on the source calibration split to meet a 5% selective-risk budget, then apply that exact threshold to each target and measure risk on the untouched target-test set. Under shift the source threshold realizes risk 0.109 (95% CI [0.098, 0.121]), more than double the budget, while coverage falls to 0.29; 71% of shifted pairs exceed the 5% budget. An oracle threshold re-tuned on a disjoint target-calibration split lowers risk to 0.077 [0.067, 0.084] at coverage 0.20 (versus the source threshold’s coverage 0.29), so re-tuning trades coverage for lower risk but, tellingly, still does not reach the 5% budget. We are careful about what this shows: it does not establish that a 5% target risk is *intrinsically* unattainable, since finite target-calibration data and empirical threshold selection themselves inflate realized risk; it shows that the source-chosen operating point transfers poorly and that even an honestly re-tuned threshold, on 250 target labels, does not fully recover the

budget. A coverage policy tuned in-domain therefore silently admits roughly twice its intended error rate under shift.

What predicts the transfer gap?

If the gap is a parameter-transfer failure, it should be predictable from how far the target’s confidence behavior has moved from the source, and in particular from how much the optimal temperature itself has moved. For every shifted source–target pair we correlate the temperature-family transfer gap with five signals, two of them genuinely *label-free* (confidence-distribution JS divergence and mean predictive-entropy shift) and three requiring labels or an oracle (optimal-temperature shift, accuracy drop, and the ATC accuracy-estimation *error*, whose computation needs the true target accuracy even though the underlying ATC estimate is label-free). We report pooled Spearman coefficients as descriptive summaries, together with the range of the coefficient computed within each seed (Table 8). The clearest predictor is the drift in the optimal temperature. Because the raw distance $|T_{\text{src}} - T_{\text{tgt}}|$ depends on the arbitrary logit scale (Section 8), we use the scale-invariant log-ratio $|\log(T_{\text{tgt}}/T_{\text{src}})|$, unchanged by a common rescaling of logits and temperatures; it has pooled $\rho =$

Table 3 Calibration transfer gap (source-fitted-temperature ECE minus target-fitted-temperature ECE), in-domain versus under shift, across the model spectrum on sentiment, all under the leakage-free split (disjoint target-calibration and target-test). The shifted gap is positive for every family, with the in-domain gap indistinguishable from zero; its magnitude diminishes for the fully fine-tuned transformer. Brackets are 95% bootstrap CIs over seeds.

Model family	Gap (in-domain)	Gap (shift)
Linear (LR/LSVM, TF-IDF)	0.003 [−0.001, 0.006]	0.021 [0.015, 0.029]
Frozen DistilBERT features	0.002 [−0.007, 0.007]	0.045 [0.036, 0.056]
Fine-tuned DistilBERT	−0.001 [−0.002, 0.001]	0.007 [0.005, 0.009]

Table 4 Calibration transfer gap (source-fitted minus target-fitted ECE, 15-bin) across three calibrator families of increasing flexibility (sentiment, five seeds, two linear model families), under the leakage-free split (disjoint target-calibration and target-test, matched N_{cal}). Brackets are 95% bootstrap CIs over seeds. The gap widens under shift for every family, showing that family inadequacy alone cannot explain the observed degradation.

Calibrator family	Gap (in-domain)	Gap (shift)
Temperature (scalar)	0.003 [−0.001, 0.006]	0.021 [0.015, 0.029]
Platt (affine logit)	0.001 [−0.001, 0.003]	0.020 [0.014, 0.025]
Isotonic (non-parametric)	0.006 [0.001, 0.009]	0.035 [0.029, 0.043]

0.75, somewhat stronger than the raw distance (pooled $\rho = 0.68$), consistent with the gap being closely associated with recalibration-parameter drift rather than with the calibrator’s form. No fitted temperature approached the optimizer bounds [0.05, 10] (observed range 0.45–2.6), so the drift is not a boundary artifact. Among purely label-free signals, the Jensen–Shannon divergence between source and target confidence distributions (pooled $\rho = 0.49$) and the shift in mean predictive entropy (pooled $\rho = 0.45$) are both moderately predictive. Notably, the ATC accuracy-estimation error is weak and inconsistent across seeds (pooled $\rho = 0.28$, per-seed range spanning zero), even though ATC estimates target accuracy reasonably (mean absolute error 0.057); this dovetails with Proposition 1, that a signal tracking target *accuracy* need not track the *calibration* gap. We report these pooled Spearman coefficients as descriptive summaries only. Because the shifted pairs within a seed share source models and target sets, a pooled-pair bootstrap would not respect that dependence, so instead of attaching pooled-pair intervals we report the coefficient computed *within each seed*: the log-ratio drift d_T is positively associated with the gap in every run (per-seed ρ range [0.47, 0.91]), as is confidence-JS ([0.31, 0.71]), so the associations are not an artifact of pooling dependent cells. We caution that the specific ranking of these predictors is itself regime-dependent (Section 6

finds the ATC accuracy-estimation error, a label-dependent diagnostic, *is* predictive on the tabular benchmark, unlike here), so we treat the temperature drift as the reference diagnostic and the genuinely label-free signals (confidence-JS, entropy shift) as regime-dependent proxies for it, and read all these as associations rather than causal claims. The same characterization explains the near-zero estimated MNLI gap of Table 2, which we quantify (Table 7): both drivers are smaller for RoBERTa across MNLI genres than for the sentiment shifts, so RoBERTa/MNLI sits in the low-drift regime where the characterization predicts, and we observe, no detectable gap.

Does the drift-gap association survive independent estimation?

Because d_T and the transfer gap are both computed from the fitted target temperature, their association could partly reflect shared estimation noise and their mathematical link rather than a property of the shift. We test this directly (Table 6). We split the source- and target-calibration sets into disjoint halves A and B (matched size 125), estimate the drift from A , and compute the transfer gap from *independently* fitted group- B temperatures on the untouched target-test set. Holding the sample size fixed at 125 per fit, the same-group association (drift and gap sharing the fitted temperatures) is $\rho = 0.64$, whereas under independent estimation it drops

Table 5 Source-fit-minus-target-fit difference (temperature family) is robust to the calibration metric: in-domain and shifted value, paired shifted-minus-in-domain increase, and 95% seed-level bootstrap CI (sentiment, five seeds, two linear models). All under the leakage-free split.

Metric	In-domain	Shift	Increase	95% CI
ECE (10 bins)	0.006	0.022	+0.017	[0.008, 0.027]
ECE (15 bins)	0.003	0.021	+0.019	[0.009, 0.030]
ECE (20 bins)	0.002	0.019	+0.017	[0.010, 0.027]
Adaptive ECE	0.002	0.015	+0.013	[0.005, 0.022]
Brier	0.000	0.003	+0.003	[0.001, 0.004]
NLL	0.002	0.013	+0.012	[0.006, 0.018]

to $\rho = 0.35$, with a per-seed range that spans zero ($[-0.08, 0.71]$), so the association is not consistently positive across the five seeds in this independent 125-example analysis. For context, the shared-fit association strengthens with calibration sample size, from $\rho = 0.64$ at 125 examples to $\rho = 0.75$ at the full 250 (Table 8), computed by the same pipeline. A no-shift reference is informative: fitting two independent same-domain source temperatures and measuring their movement $|\log(T_A/T_B)|$ gives a mean of 0.227 from sampling alone, against a mean cross-domain drift of 0.385 at the same sample size; we report these as comparable magnitudes rather than decomposing one into the other. We therefore read the drift diagnostic conservatively: it is a retrospective association that weakens substantially under independent calibration estimation and, at 125 examples per fit, is inconsistent across seeds. This qualifies the diagnostic; the transfer gap itself is measured on the full disjoint splits and its 250-example estimate remains positive (Table 1), though that estimate is likewise subject to calibration-sample variability.

A label-free indicator of severe transfer failure.

Correlation suggests, but does not establish, that a label-free signal can operationally flag calibration drift. We test this directly as a detection task: label a shifted pair *severe* if its (leakage-free) transfer gap exceeds a threshold τ_Δ , and rank pairs by a label-free signal. At $\tau_\Delta = 0.03$ (prevalence 0.30 over 120 experiment observations, i.e. 12 directional domain pairs $\times$ 2 models $\times$ 5 seeds, not 120 distinct shifts), confidence-JS divergence detects severe failures with pooled AUROC 0.81 and average precision 0.75; entropy

Table 6 Independent-estimation test of the drift-gap association (linear sentiment, shifted pairs, five seeds). At a fixed 125 examples per fit, estimating the drift and the gap from disjoint calibration halves (A and B) roughly halves the pooled association relative to the same-group estimate and yields a per-seed range spanning zero; the shared-fit association is stronger at the full 250-example size (Table 8, same pipeline). The no-shift reference shows temperature movement of comparable magnitude from sampling alone.

Quantity	Value
Same-group, size 125 (ρ)	0.64
Independent A/B , size 125 (ρ)	0.35
(per-seed range)	$[-0.08, 0.71]$
Same-group, full 250 (Table 8)	0.75
No-shift movement $ \log(T_A/T_B) $ (mean)	0.227
Cross-domain d_T^A , size 125 (mean)	0.385

shift is comparable (AUROC 0.79). The ranking is insensitive to the definition of “severe”: AUROC for confidence-JS is 0.76, 0.81, 0.81, 0.86 at $\tau_\Delta \in \{0.02, 0.025, 0.03, 0.04\}$. We stress that this is a *pooled, in-sample* ranking; its out-of-sample behaviour is weaker (next paragraph), so we present it as an exploratory indicator for prioritizing recalibration or target-label collection, not as a validated classifier.

Out-of-sample and cross-regime validation.

To rule out the concern that the indicator is characterized and evaluated on the same shifts, we validate it out-of-sample. In a leave-one-target-domain-out protocol, we select a confidence-JS decision threshold τ_{JS} (severe iff confidence-JS $> \tau_{JS}$) on the training domains by maximizing Youden’s J (sensitivity + specificity -1), then evaluate on the held-out domain, with “severe” defined by $\tau_\Delta = 0.03$ as above. Because the folds have very unequal severe prevalence, we report

Table 7 Quantitative contrast underlying the regime dependence (leakage-free split). The unlabeled confidence-distribution divergence and the scale-invariant parameter drift $d_T = |\log(T_{\text{tgt}}/T_{\text{src}})|$ (with the scale-dependent raw distance shown for reference) are both larger for the sentiment shifts than for RoBERTa across MNLI genres, matching the larger transfer gap. Means over all shifted pairs; the transfer gap column reports 95% bootstrap CIs over seeds.

Setting	Confidence-JS	d_T (log-ratio)	raw $ T_{\text{src}} - T_{\text{tgt}} $	Transfer gap (shift)
Sentiment (linear)	0.065	0.295	0.279	0.021 [0.015, 0.029]
MNLI (RoBERTa)	0.029	0.118	0.135	−0.000 [−0.005, 0.003]

Table 8 What predicts the temperature-family transfer gap (leakage-free split): pooled Spearman coefficient over all shifted source–target pairs (sentiment, five seeds, two linear model families), reported as a descriptive summary, together with the range of the coefficient computed *within each seed* (a dependency-preserving check, since pooled pairs share source models and target sets). We do not attach pooled-pair confidence intervals, which would not respect that dependence. The scale-invariant optimal-temperature drift is the strongest and most consistent predictor; the ATC accuracy-estimation error is weak and inconsistent across seeds, consistent with Proposition 1. [†]The ATC accuracy *estimate* is label-free, but its *error*, the signal correlated here, requires the true target accuracy and so is not label-free.

Signal	Label-free?	Pooled ρ	Per-seed ρ range
$ \log(T_{\text{tgt}}/T_{\text{src}}) $ (scale-invariant drift)	no (diagnostic)	0.75	[0.47, 0.91]
raw $ T_{\text{src}} - T_{\text{tgt}} $ (scale-dependent)	no (diagnostic)	0.68	[0.39, 0.80]
Confidence-distribution JS divergence	yes	0.49	[0.31, 0.71]
Accuracy drop	no	0.55	[0.38, 0.69]
Mean predictive-entropy shift	yes	0.45	[0.40, 0.53]
ATC accuracy-estimation error	no [†]	0.28	[−0.19, 0.65]

balanced accuracy rather than raw accuracy. The *Yelp* held-out fold contains no severe pairs at all (prevalence 0), so sensitivity, AUROC, and balanced accuracy are undefined there; we therefore exclude it from the averages and average over the three evaluable folds, obtaining mean held-out balanced accuracy 0.62. Of those, RottenTomatoes and SST-2 reach balanced accuracy 0.70 and 0.61 (fold AUROC 0.83 and 0.66), while the low-prevalence IMDB fold is near chance; the signal is thus useful but far from a reliable classifier out-of-sample, and we report this honestly rather than as a headline. More stringently, we apply the indicator *across tasks*: a threshold fit only on the sentiment shifts is applied to the RoBERTa/MNLI setting it never saw, and correctly places RoBERTa/MNLI (low confidence-JS) in the *not-severe* regime, matching the observed near-zero gap. The unlabeled divergence signal, developed on one task, therefore correctly places a different task and model in the low-transfer-gap regime, even though its absolute decision threshold does not port across modalities (Section 6).

5 The transfer gap grows with shift severity

The sentiment-versus-NLI contrast shows two regimes; a controlled dose-response lets us observe how the associated quantities move together as shift severity increases. We train on clean IMDB and evaluate on the same domain under graded word-level corruption (deletion, adjacent swap, substitution) at severity $\delta \in \{0, \dots, 5\}$, measuring at each level the accuracy, confidence-JS from the source calibration split, optimal-temperature drift $|T_{\text{src}} - T_{\text{tgt}}|$, and transfer gap (five seeds; Table 9). We run this leakage-free: at each severity the corrupted target set is split into a disjoint target-calibration subset (fitting T_{tgt}) and an untouched target-test subset (evaluating every ECE). As severity rises, accuracy falls monotonically ($0.86 \rightarrow 0.79$) and both the parameter drift and the gap rise together: the gap grows from -0.002 at $\delta=0$ to 0.023 at $\delta=5$ (Table 9). Because the six severity levels are repeated measurements within each of five runs, we report the trend *per seed* rather than pooling the 30 dependent observations: the within-seed Spearman correlation of severity with the gap is positive in every

run (mean $\rho = 0.77$, range $[0.54, 1.00]$ across the five seeds), as is severity $\rightarrow$ parameter-drift (mean 0.74) and parameter-drift $\rightarrow$ gap (mean 0.82). The confidence-JS $\rightarrow$ gap link is weaker and less consistent (mean 0.58, range $[-0.03, 0.89]$), which we report rather than overstate. The hypothesized pathway *shift severity $\rightarrow$ confidence-distribution drift $\rightarrow$ recalibration-parameter drift $\rightarrow$ transfer gap* is thus supported at the severity and parameter-drift links and only suggestive at the confidence-JS link within this single controlled setting. As correlational evidence within one task, this is consistent with, though it does not prove, a causal pathway, and it complements the cross-dataset contrast of Table 7.

6 External validation: natural shift on a tabular benchmark

The evidence so far is confined to text. Two of its pillars, the sentiment/NLI contrast and the controlled dose-response, could in principle be artifacts of an NLP pipeline or of synthetic corruption. To test whether the transfer gap and its mechanism survive a move to a different modality and a genuinely *natural* shift, we replicate the study on an established tabular benchmark, **ACSIncome** from Folktables (Ding et al. 2021): predict whether a US-census respondent earns above \$50k, with US *states* as domains. States share the feature and label space but differ naturally in covariate and label-conditional distribution, so cross-state evaluation is a natural domain shift rather than an induced corruption. *Domain-selection rule.* We fixed the domain set before computing any transfer-gap outcome, by a predetermined, outcome-independent criterion: the four largest states by ACS sample size (CA, TX, NY, FL) plus Puerto Rico, included *a priori* as an expected strong demographic and socioeconomic shift, not because it produced a large gap. We use standardized numeric features with ℓ_2 logistic regression as the linear-model analog, and five seeds. Crucially, we run this through the *identical* protocol: the metric, temperature/Platt/isotonic fitting, confidence-JS, $|T_{\text{src}} - T_{\text{tgt}}|$, and transfer-gap code are imported unchanged from the sentiment experiment (Appendix A); only the data source and feature extractor differ.

One benchmark, both regimes, ordered by parameter drift.

The tabular benchmark does not merely reproduce a single number; it recreates *both* regimes of the paper within one setting, and the optimal-temperature drift orders them (Table 10). Mean target accuracy falls from 0.801 in-domain to 0.747 under state shift, a drop of 0.054 relative to the in-domain diagonal (equivalently, the mean per-pair drop against the diagonal baseline is 0.054). Among mainland-only shifts (CA, TX, NY, FL), the confidence-space shift is mild (scale-invariant $d_T = 0.16$, raw $|T_{\text{src}} - T_{\text{tgt}}| = 0.18$, accuracy drop 0.011) and the temperature transfer gap stays small (0.003, at the in-domain level 0.003), the low-drift/low-gap regime of RoBERTa/MNLI. Shifts involving Puerto Rico induce a large confidence-space shift ($d_T = 0.66$, raw 1.06, accuracy drop 0.119) and a correspondingly large gap (0.059, exceeding the sentiment magnitude), the high-gap regime. This ordering is exactly what the characterization predicts, and across *all* shifted state pairs the scale-invariant parameter drift $|\log(T_{\text{tgt}}/T_{\text{src}})|$ is associated with the gap at $\rho = 0.63$, positive in both settings (leakage-free sentiment $\rho = 0.75$), and holds across the Platt and isotonic families too (with isotonic again the largest). That an independent tabular benchmark with natural geographic shift reproduces both regimes and the same parameter-drift ordering provides evidence that the gap and its association with parameter drift are not specific to text data or synthetic corruption. We stress that no single benchmark-wide gap magnitude is a meaningful target: the gap tracks the size of the confidence-space shift, which is precisely why $|T_{\text{src}} - T_{\text{tgt}}|$, not a fixed number, is the diagnostic.

An honest boundary: the warning signal transfers in rank but not in scale.

The confidence-divergence signal transfers only partially, and we report the limit rather than smooth it over. In *relative* terms it remains useful: on the tabular pairs, confidence-JS detects severe transfer failure (gap > 0.03) with threshold-free AUROC 0.77. But the *absolute* sentiment-derived decision threshold does *not* port: tabular confidence divergences occupy a much smaller numeric range (mean 0.015, max 0.040) than sentiment (mean 0.065), so a frozen sentiment threshold

Table 9 Controlled shift-severity dose-response (train on clean IMDB, evaluate under graded corruption; means over five seeds), under the leakage-free split (target temperature fit on a disjoint target-calibration subset, gap scored on the untouched target-test subset). Accuracy falls while optimal-temperature drift and the transfer gap rise together; confidence-JS rises but is a weaker correlate under the honest split.

Severity δ	Accuracy	Confidence-JS	$ T_{\text{src}} - T_{\text{tgt}} $	Transfer gap
0	0.857	0.004	0.039	-0.002
1	0.843	0.005	0.063	-0.001
2	0.832	0.005	0.072	0.003
3	0.821	0.006	0.069	0.010
4	0.800	0.007	0.146	0.025
5	0.791	0.008	0.215	0.023

never fires and its out-of-the-box sensitivity is zero. The divergence signal’s rank-ordering of shifts is therefore portable across modalities, but its absolute scale is not; a deployable warning rule must be recalibrated per modality. Relatedly, the ATC accuracy-estimation error, a label-dependent diagnostic (its computation needs the true target accuracy), was weak and inconsistent about the gap on sentiment ($\rho = 0.28$) yet *is* predictive here ($\rho = 0.52$); this regime difference does not contradict Proposition 1, which concerns what mean-level statistics can *guarantee*, not what they happen to correlate with in a given benchmark, but it shows that even a label-dependent proxy’s informativeness is itself regime-dependent. Taken together, one quantity, the scale-invariant optimal-temperature drift $d_T = |\log(T_{\text{tgt}}/T_{\text{src}})|$, is associated with the gap consistently across both text and tabular data ($\rho = 0.75$ and 0.63), whereas the deployable label-free proxy (confidence-JS) transfers only in rank and not absolute scale, and the label-dependent ATC-error diagnostic changes informativeness across regimes: parameter drift appears to be the more consistent reference, with the proxies varying by regime.

7 Limits of confidence-based label-free recalibration

Given the transfer gap, we ask whether label-free recalibration closes it. We first show a structural reason it is hard, then evaluate the strategies empirically.

7.1 Why mean-level signals cannot identify calibration

The natural label-free strategies rescale confidence to match a target statistic: mean-matching sets the mean confidence equal to a label-free accuracy estimate (e.g. ATC (Garg et al. 2022)); variance-matching matches the confidence spread. The following shows such statistics do not identify the conditional-correctness curve, and hence not the calibration error, so no rule chosen solely from these statistics can be guaranteed to recover a calibrated map that preserves informative confidence variation (a trivial calibrated-but-uninformative map may still exist, as Remark 1 notes).

We state the results in terms of the *population calibration error*

$$\text{CE} = \mathbb{E}[|\Pr(Y = 1 | S) - S|], \quad (4)$$

which does not depend on a bin count. The constructions below also produce a discrepancy in binned ECE whenever the distinct confidence support points occupy distinct bins.

Proposition 1 (Mean confidence and accuracy do not identify calibration) *There exist two joint laws over confidence $S \in [0, 1]$ and correctness $Y \in \{0, 1\}$ with identical mean confidence $\mathbb{E}[S]$ and identical accuracy $\mathbb{E}[Y] = \Pr(Y = 1)$ but different calibration error CE. Hence mean confidence and target accuracy do not determine the calibration error.*

Proof Fix $\mathbb{E}[S] = 0.7$ and $\mathbb{E}[Y] = 0.7$. Law $\mathcal{P}_A$: all mass at $S = 0.7$ with $Y \sim \text{Bernoulli}(0.7)$ independent of S ; then $\Pr(Y = 1 | S = 0.7) = 0.7 = S$, so $\text{CE}(\mathcal{P}_A) = 0$. Law $\mathcal{P}_B$: half the mass at $S = 0.9$ with $\Pr(Y = 1 | S = 0.9) = 0.5$, half at $S = 0.5$ with

Table 10 External replication on ACSIncome (Folktables), US states as natural-shift domains, five seeds, standardized-feature logistic regression, run through the identical protocol. A single benchmark spans both regimes: mainland-only shifts are mild and the gap stays near its in-domain value, whereas Puerto-Rico-involving shifts are strong and the gap is large. Across *all* shifted pairs the scale-invariant parameter drift $|\log(T_{\text{tgt}}/T_{\text{src}})|$ is associated with the gap at $\rho = 0.63$ (leakage-free sentiment: 0.75), ordering the two regimes. All numbers use the leakage-free split (disjoint target-calibration and target-test); brackets are 95% bootstrap CIs over seeds.

Transfer gap	In-domain	Mainland shift	PR-involving shift
Temperature	0.003 [0.000, 0.005]	0.003 [−0.002, 0.007]	0.059 [0.052, 0.066]
Platt	0.004 [−0.001, 0.009]	0.003 [−0.003, 0.008]	0.106 [0.088, 0.120]
Isotonic	0.001 [−0.002, 0.004]	0.024 [0.018, 0.029]	0.159 [0.140, 0.177]
$d_T = \log(T_{\text{tgt}}/T_{\text{src}}) $ (mean)	—	0.16	0.66
raw $ T_{\text{src}} - T_{\text{tgt}} $ (mean)	—	0.18	1.06
Accuracy drop vs. diagonal (mean)	—	0.011	0.119
<i>Across all shifted pairs: scale-invariant $d_T \rightarrow$ gap Spearman $\rho = 0.63$ (sentiment: 0.75);</i>			
<i>ATC accuracy-estimation error $\rightarrow$ gap $\rho = 0.52$ (sentiment: 0.28), reversing the ranking.</i>			

$\Pr(Y = 1 \mid S = 0.5) = 0.9$. Then $\mathbb{E}[S] = \frac{1}{2}(0.9+0.5) = 0.7$ and $\mathbb{E}[Y] = \frac{1}{2}(0.5+0.9) = 0.7$, matching $\mathcal{P}_A$, while $\text{CE}(\mathcal{P}_B) = \frac{1}{2}|0.9 - 0.5| + \frac{1}{2}|0.5 - 0.9| = 0.4$. The two laws agree on both scalars yet differ in calibration, so these scalars do not determine CE. $\square$

Proposition 2 (Variance does not identify calibration either) *There exist two laws with identical mean confidence, identical accuracy, and identical confidence variance, but different calibration error.*

Proof Keep $\mathcal{P}_B$ from above: confidence takes values $\{0.5, 0.9\}$ each with probability $\frac{1}{2}$, so $\mathbb{E}[S] = 0.7$, $\text{Var}(S) = 0.04$, $\mathbb{E}[Y] = 0.7$, $\text{CE} = 0.4$. Construct $\mathcal{P}_C$ on the same confidence support with the same marginal (values $\{0.5, 0.9\}$ each with probability $\frac{1}{2}$, giving identical $\mathbb{E}[S] = 0.7$ and $\text{Var}(S) = 0.04$) but calibrated conditionals $\Pr(Y = 1 \mid S = 0.9) = 0.9$ and $\Pr(Y = 1 \mid S = 0.5) = 0.5$. Then $\mathbb{E}[Y] = \frac{1}{2}(0.9 + 0.5) = 0.7$ matches, yet $\text{CE}(\mathcal{P}_C) = 0$. The two laws share mean confidence, variance, and accuracy but have different calibration error, so variance-matching is unidentified as well. $\square$

Population calibration error is a functional of the full conditional-accuracy relationship $p \mapsto \Pr(Y = 1 \mid S = p)$, whereas ATC and mean/variance matching constrain only low-order moments of S and its aggregate accuracy; empirical ECE approximates the corresponding calibration discrepancy. A correct target accuracy estimate is therefore insufficient on its own: it pins down where confidence should sit *on average*, not the *shape* of the miscalibration that ECE

penalizes. This also illustrates why the pseudo-ECE objective remains underidentified: a single threshold-derived proxy for correctness cannot in general recover the full conditional-correctness curve it is trying to match.

Remark 1 (Scope of the impossibility) Propositions 1–2 state that mean confidence, accuracy, and variance do not identify the *conditional-calibration curve* $p \mapsto \Pr(Y = 1 \mid S = p)$, hence not CE or ECE. They do *not* claim that no calibrated output can be produced from these quantities: given exact target accuracy α , the constant map $g(S) \equiv \alpha$ is trivially calibrated ($\text{CE} = 0$), but it discards all information about which inputs are easier or harder, removing the informative score ordering selective prediction needs (different coverages remain reachable only through arbitrary tie-breaking). The meaningful claim is therefore that mean-level statistics cannot recover a calibration map that *preserves useful confidence variation*, i.e. that both attains low CE and retains discrimination; scalar rescaling to match a moment cannot do this in general.

Proposition 3 (A positive companion: when mean-matching succeeds) *Let S be the binary maximum confidence, so $S \in [\frac{1}{2}, 1)$ with $\Pr(S > \frac{1}{2}) > 0$, and write the (nonnegative) margin $M = \sigma^{-1}(S) \geq 0$; temperature scaling maps S to $g_T(S) = \sigma(M/T)$. Suppose the target conditional-correctness curve is exactly representable in this family, i.e. there is a $T^* > 0$ with $\Pr(Y = 1 \mid S) = \sigma(M/T^*)$. Then $\mathbb{E}[\sigma(M/T)]$ is strictly decreasing in T , so mean-matching to the target accuracy α has a unique solution, and it equals T^* . Under this correct specification, mean-matching*

with an exact accuracy estimate recovers the target calibration map.

Proof sketch For each point with $M > 0$, $g_T(S) = \sigma(M/T)$ is strictly decreasing in T (toward $\frac{1}{2}$); on the event $\{M > 0\}$, which has positive probability by assumption, this makes $\mathbb{E}[\sigma(M/T)]$ continuous and strictly decreasing in T , so the level set $\{T : \mathbb{E}[\sigma(M/T)] = \alpha\}$ is a singleton. Well-specification gives $\mathbb{E}[\sigma(M/T^*)] = \mathbb{E}[\Pr(Y = 1 \mid S)] = \alpha$, so the unique solution is T^* . $\square$

We stress that the nonnegative-margin structure is essential: the same monotonicity fails for an arbitrary class probability $S \in [0, 1]$. For instance, if S takes values 0.25 and 0.75 with equal probability, symmetry forces $\mathbb{E}[\sigma(\sigma^{-1}(S)/T)] = \frac{1}{2}$ for every T , so mean confidence carries no information about T . Proposition 3 is therefore a *sufficient condition for success*, not a theorem about failure: it establishes when mean-matching provably works (correct specification within the temperature family on the max-confidence scale), and outside its assumptions it makes no prediction, leaving what actually happens to the experiments. Propositions 1–3 together reframe the question from “does mean-matching work?” to “*when* does it work?”.

7.2 Empirical confirmation

Tested with the linear models across five seeds under the leakage-free protocol (source temperature and ATC threshold fit on the source-calibration split, oracle on the disjoint target-calibration split, all ECE evaluated on the untouched target-test set), none of the three strategies beats the source calibrator; this is the behaviour the identifiability limit permits, and Proposition 3 gives only a sufficient condition for success (correct specification within the temperature family) that we have no reason to expect holds under shift. Mean-matching is in fact *worse* than the source temperature on sentiment (mean shifted ECE 0.064 vs. 0.054; seed-level paired difference +0.010, 95% bootstrap CI over the five seeds [+0.001, +0.021], excluding zero), even though the ATC accuracy estimate is reasonable (mean absolute error 0.057 against true target accuracy): matching an approximate mean does

not yield calibration. The distributional pseudo-ECE objective is far worse (+0.130 over source-T, CI [+0.114, +0.145]) and variance-matching is also worse (+0.008, CI [+0.003, +0.014]). We report seed-level effect sizes with bootstrap CIs rather than a rank-test p -value: a two-sided exact Wilcoxon signed-rank test on five nonzero paired differences cannot attain $p < 2/2^5 = 0.0625$, so we summarise the effect with a CI over seeds instead. Table 11 summarizes; only the target-fitted temperature, refit on disjoint target labels, improves calibration (0.032, difference −0.021 CI [−0.029, −0.015]).

On a fine-tuned transformer, and against a CPCS-inspired baseline.

Two objections remain: the heuristics were only tested on linear models, and simple hand-designed rules are a weak target. We address both by rerunning the label-free strategies on *fine-tuned DistilBERT* (sentiment, three seeds, all shifted source–target pairs) and adding a CPCS-inspired importance-weighted temperature-scaling baseline (Park et al. 2020), which reweights the source calibration set by an estimated target/source density ratio before fitting temperature (Table 12). This tests whether unlabeled density-ratio reweighting, in the spirit of CPCS, closes the gap; it is a lightweight adaptation (Appendix A) and is not intended as a faithful reproduction of CPCS. We run this under the *same* leakage-free protocol as the model-spectrum result (Table 3) and, crucially, compute every interval with the *same* seed-level paired-difference bootstrap, so the two tables are numerically consistent: the oracle-versus-source difference here is −0.007 (CI [−0.009, −0.005]), the exact sign-reversal of the source-versus-target gap +0.007 ([0.005, 0.009]) in Table 3 (source-T 0.054, oracle 0.046). The conclusion holds on the transformer: mean-matching (0.095) and pseudo-ECE (0.220) are significantly worse than the source temperature (0.054), and variance-matching (0.058) is no better. The CPCS-inspired baseline is statistically indistinguishable from the source temperature (paired difference +0.001, 95% CI [−0.000, +0.003]): reweighting by this unlabeled density ratio does not recover the target’s confidence–accuracy relationship. Only the target-fitted temperature improves calibration,

Table 11 Mean shifted-domain ECE for three label-free recalibration strategies against the source-fitted and target-fitted temperatures (sentiment, five seeds, two linear model families), all under the leakage-free split (oracle fit on a disjoint target-calibration set, every ECE scored on the untouched target-test set). No label-free strategy beats the source temperature; only the target-fitted temperature improves calibration. The “vs. Source-T” column is the seed-level paired difference with a 95% bootstrap CI over the five seeds.

Strategy	Shifted ECE	vs. Source-T [95% CI]
Source temperature (T_{src})	0.054	—
Mean-matching (ATC)	0.064	+0.010 [+0.001, +0.021]
Pseudo-ECE (distributional)	0.184	+0.130 [+0.114, +0.145]
Variance-matching	0.062	+0.008 [+0.003, +0.014]
Oracle (T_{tgt} , uses labels)	0.032	−0.021 [−0.029, −0.015]

by the small 0.007 consistent with the modest fine-tuned-model gap. None of the confidence-based label-free strategies we tested, including the confidence-only density-ratio reweighting, improves on the source calibrator; effective label-free recalibration under shift therefore appears to need target information beyond these mean-level and confidence-only signals. We do not claim this about unlabeled reweighting in general: a richer density-ratio estimator over representations is untested and could behave differently.

A stronger, feature-based importance-weighting baseline.

A reasonable objection is that the CPCS-inspired baseline above fails only because it uses scalar confidence as its sole domain-classifier feature. We test this on the tabular benchmark, where the original ten ACSIncome features are available, by fitting the domain classifier on those features and reweighting the source-calibration loss by its (cross-fitted, a-priori-clipped) density ratios; weights are estimated without ever consulting target evaluation labels (Appendix A). This feature-based implementation produces little change in mean ECE relative to the source calibrator (Table 13); we attribute the result to this implementation as a whole (feature input, cross-fitting, and percentile clipping all differ from the confidence-only variant), not to the feature input alone. Feature-based weighting reaches mean shifted ECE 0.0717 versus source temperature 0.0713, a seed-level paired difference of +0.0004 (95% CI [−0.0023, +0.0030], including zero), and confidence-only weighting is similar (0.0714, +0.0001 [−0.0002, +0.0005]); ECE improved in 51 of the 100 pair-seed evaluations (20 directed state pairs $\times$ five seeds), which we report descriptively

rather than against a 50% chance benchmark, since these evaluations are not independent shifts. On NLL the feature-based method shows a small mean reduction relative to the source temperature (0.532 vs. 0.536; paired difference −0.0042, CI [−0.0092, +0.0007], including zero), so we do not claim a detectable NLL benefit over the source calibrator; we do not test the feature-versus-confidence-only NLL difference. By contrast the target-label oracle clearly improves both metrics (ECE −0.025, CI [−0.029, −0.020]; NLL −0.043, CI [−0.050, −0.035]). The weights are not highly concentrated on average (mean normalized effective sample size $\text{ESS}/n = 0.74$, about 185 of 250 effective observations; mean max-to-mean weight 3.5), though this does not rule out that weight instability or inaccurate density ratios contribute in individual pairs. We qualify the finding: the ACS state shifts are not established to be pure covariate shift, and under label-conditional change even a correct density ratio need not recover calibration, so this bounds the tested feature-based variant rather than importance weighting in general.

7.3 How much target supervision is needed?

If label-free recalibration cannot close the gap but a target-fitted temperature can, the actionable question is *how many* target labels are required. We answer it directly with a label-budget sweep: on the shifted sentiment pairs we refit the target temperature on $B \in \{25, 50, 100, 250\}$ labelled target-calibration examples (disjoint from the target-test set) and measure how much of the source-temperature ECE it recovers (Table 14). At only 25 labels the refit does not

Table 12 Label-free recalibration on fine-tuned DistilBERT (sentiment, three seeds, shifted pairs), under the leakage-free split (oracle on a disjoint target-calibration set; ECE on the untouched target-test set), including a CPCS-inspired importance-weighted baseline (Park et al. 2020). Mean shifted ECE with 95% seed-level bootstrap CIs, and the paired difference versus the source temperature computed with the identical seed-level procedure. The oracle-vs-source difference ($-0.007 [-0.009, -0.005]$) is the exact sign-reversal of the source-vs-target gap in Table 3. No label-free method beats the source temperature; only the label-using oracle does.

Strategy	Shifted ECE [95% CI]	vs. Source-T
Source temperature	0.054 [0.051, 0.058]	—
Mean-matching (ATC)	0.095 [0.080, 0.107]	+0.041 [+0.029, +0.056]
Pseudo-ECE (distributional)	0.220 [0.199, 0.232]	+0.166 [+0.140, +0.181]
Variance-matching	0.058 [0.055, 0.060]	+0.005 [−0.003, +0.009]
CPCS-inspired (importance-wtd.)	0.055 [0.053, 0.058]	+0.001 [−0.000, +0.003]
Oracle (T_{tgt})	0.046 [0.042, 0.051]	−0.007 [−0.009, −0.005]

Table 13 Feature-based importance-weighted temperature calibration on ACSIncome (five seeds, shifted state pairs), under the leakage-free split. A domain classifier on the original ten features (cross-fitted, a-priori clip) produces little change in mean ECE and a small NLL reduction relative to the source temperature whose paired interval includes zero; only the target-label oracle detectably improves both metrics. Paired intervals are versus the source temperature. ECE and NLL are means over pair-seed evaluations; “vs. src” is the seed-level paired difference with a 95% bootstrap CI over the five seeds; “pairs improved” counts the 100 pair-seed evaluations (20 directed pairs $\times$ five seeds) on which ECE beats source-T (reported descriptively, not against a chance benchmark).

Strategy	ECE	ECE vs. src [95% CI]	NLL	Pairs impr.
Source temperature	0.0713	—	0.536	—
Confidence-only weighting	0.0714	+0.0001 [−0.0002, +0.0005]	0.536	50/100
Feature-based weighting (10 feats)	0.0717	+0.0004 [−0.0023, +0.0030]	0.532	51/100
Oracle (T_{tgt} , labels)	0.0461	−0.0252 [−0.0288, −0.0201]	0.494	—

improve on the source temperature (mean change -0.008 , CI $[-0.020, +0.004]$, which includes zero, so we do not claim statistically established harm, only no detectable average benefit); the point estimate is negative, consistent with the temperature being estimated too noisily. From 50 labels onward the mean improvement is positive and its CI excludes zero, rising with budget: $+0.011$ ($[+0.005, +0.016]$) at 50, $+0.012$ ($[+0.002, +0.021]$) at 100, and $+0.021$ ($[+0.015, +0.029]$) at 250. Stated against the 250-label reference gap of 0.021, this is a recovery of roughly 52% at 50 labels and 57% at 100. The supported finding is therefore narrow and specific: in these linear sentiment experiments, 50–100 labelled target examples recover roughly half of the average improvement that the 250-label reference achieves. We do not claim this establishes perfect calibration (the 250-label point closes “all” of the gap only by definition of that reference) or a general sample-complexity result; the numbers are averages over shifted pairs. Examining per-pair reliability directly, the fraction of the 12 shifted

domain pairs whose ECE improves rises from 33% at 25 labels to 67% at 50 and 100 and 83% at 250: a modest budget of 50–100 labels is helpful on a majority of the evaluated domain pairs but not yet uniformly.

7.4 Keep or refit? A finite-budget decision, and its bias–variance tradeoff

The budget sweep exposes a practical decision the retrospective gap does not: given only B target labels, should a practitioner *keep* the source calibrator or *refit* it? We study this on the linear models (LR on sentiment, standardized-feature LR on ACS); for each shifted pair and budget we draw one random size- B subsample of the target-calibration split, and *all* strategies see the identical subsample. Split-CV deploys the temperature fitted on its 70% training portion (it does not refit on all B after selection), so it trades data for a validation signal. We set the decision threshold and deterioration margin on sentiment

Table 14 Target-label-budget sweep (temperature family, shifted sentiment pairs, five seeds). “Gap closed” is the source-temperature ECE minus the budget-refit ECE on the untouched target-test set, with 95% bootstrap CIs over seeds; the last column is the fraction of the 12 shifted domain pairs that improve. Relative to the 250-label reference gap (0.021), 50–100 labels recover roughly half; 25 labels show no detectable average benefit (CI includes zero).

Target labels B	ECE (src-fit)	ECE (B -refit)	Gap closed [95% CI]	% pairs improved
25	0.054	0.062	−0.008 [−0.020, +0.004]	33%
50	0.054	0.042	+0.011 [+0.005, +0.016]	67%
100	0.054	0.042	+0.012 [+0.002, +0.021]	67%
250	0.054	0.032	+0.021 [+0.015, +0.029]	83%

and apply them to ACS; because ACS has been examined elsewhere in this study, we treat it as an out-of-development-domain check for the *decision rule* rather than a benchmark seen for the first time. The naive answer, “always refit,” is actively harmful at small budgets (Table 15): at $B = 25$ it worsens target NLL relative to keeping the source on 45% of sentiment pair–seed cases and 57% of ACS cases, with a negative mean. This is the finite-sample estimation cost of refitting, and it motivates an uncertainty-aware decision.

A bias–variance criterion.

Fitting temperature by NLL, we work in the inverse temperature $\beta = 1/T$; the per-example loss has gradient $g_i = z_i(\sigma(\beta z_i) - y_i)$ and curvature $h_i = z_i^2 \sigma(\beta z_i)(1 - \sigma(\beta z_i))$ for logit z_i . Standard M -estimation theory gives the expected out-of-sample NLL improvement of refitting over the fixed source parameter as approximately

$$\mathbb{E}[\text{improvement}] \approx D - \frac{J}{2BH}, \quad (5)$$

where $D = R(\beta_{\text{src}}) - R(\beta^*)$, R is target NLL, β^* its minimiser, $H = \mathbb{E}[h]$ the curvature, and $J = \text{Var}(g)$ the gradient variance at the optimum (derivation and assumptions in Appendix A). Because the in-sample improvement $\hat{D}$ overestimates D by the same order, an estimable, bias-corrected criterion computed *entirely from the B labels* is $\hat{C} = \hat{D} - \hat{J}/(B\hat{H})$, and the rule REFIT-IF $\hat{C} > 0$. This is a direct application of existing estimation theory, not a new estimator; our question is whether it, or simpler uncertainty-aware baselines, predicts useful refitting across shifts. All labels used for fitting, validation, and the decision count toward B .

The decision, evaluated on held-out domains.

We compare five budget-respecting strategies, keep, always-refit, shrink-to-source ($\beta = \lambda\hat{\beta} + (1 - \lambda)\beta_{\text{src}}$ with $\lambda = \hat{D}^+ / (\hat{D}^+ + \hat{J}/(B\hat{H}))$), split-CV keep-vs-refit, and our $\hat{C}$ rule, freezing all design choices (threshold 0, deterioration margin $\delta=0.005$) on sentiment and applying them unchanged to ACS. The finding is a qualified, partly negative one (Table 15, Fig. 1). First, the evaluated rules *reduce* the frequency of harmful updates relative to unconditional refitting but do not eliminate it: at $B = 25$ they cut the harmful-update rate from 0.45–0.57 (always-refit) to 0.14–0.30, yet residual harm remains, and *none* achieves a positive mean NLL improvement on sentiment at 25 labels (shrink −0.020 [−0.047, −0.003], CV −0.021 [−0.046, −0.001], $\hat{C}$ −0.030 [−0.055, −0.007]; all CIs below zero). Small-budget adaptation can thus still underperform simply keeping the source calibrator. Second, our $\hat{C}$ criterion does not outperform the simpler baselines and is in fact slightly worse in paired terms: the seed-level paired difference $\hat{C}$ minus shrinkage is −0.010 [−0.021, −0.000] at $B = 25$ and −0.006 [−0.008, −0.004] at $B = 50$ on sentiment (similar on ACS), so shrinkage and CV are, if anything, the better of the evaluated rules. At the larger budgets the strategies have similar mean gains (e.g. ACS $B = 50$: $\hat{C}$ +0.034 [+0.024, +0.044], shrinkage +0.036 [+0.028, +0.045], CV +0.036 [+0.027, +0.044]). We therefore present this as a *characterization* of the keep-or-refit tradeoff with a meaningful negative result, not a superior method: refitting a scalar temperature on a few dozen target labels should be gated by an uncertainty-aware rule (shrinkage and cross-validation being the strongest here), and even then can fail to

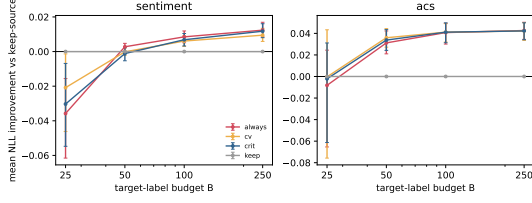


Fig. 1 Mean target-NLL improvement over keeping the source calibrator versus target-label budget B , for sentiment (decision-rule development domains) and ACS (out-of-development check). “Always refit” is harmful at small B (below zero); the uncertainty-aware rules reduce but do not eliminate the harm at $B=25$ and converge to a similar benefit as B grows. Error bars are 95% seed-level bootstrap CIs.

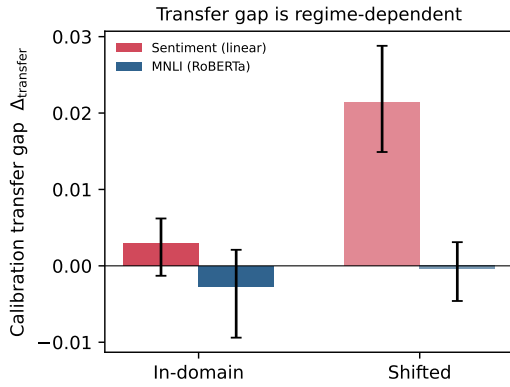


Fig. 2 Calibration transfer gap in-domain versus under shift, with 95% confidence intervals, under the leakage-free split. The gap is regime-dependent: for sentiment (linear models) it widens significantly under shift ($0.003 \rightarrow 0.021$), whereas for cross-genre MNLI with a competent fine-tuned RoBERTa it is indistinguishable from zero in both conditions. This contrast, not a universal widening, is the paper’s central empirical finding.

beat keeping the source at the smallest budgets. Decision-specific ECE and Brier move consistently with NLL (Table 16, Appendix).

8 Discussion

Parameter transfer is a substantial, but regime-dependent, source of calibration failure.

Reporting a low in-domain ECE, often after temperature scaling, as evidence of trustworthy confidence is common. On sentiment-domain shift this exact quantity degrades, for linear models, frozen

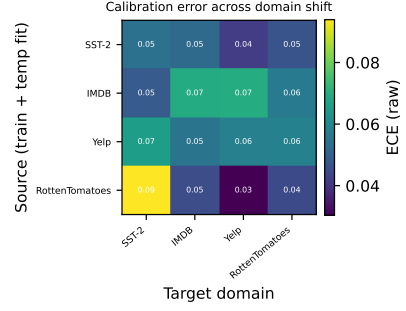


Fig. 3 Sentiment: uncalibrated ECE for every source–target pair (mean over model families). Off-diagonal (shifted) cells are typically, though not uniformly, worse than the in-domain diagonal (mean 0.053 shifted vs. 0.049 in-domain).

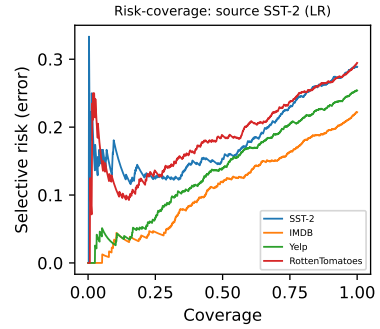


Fig. 4 Risk–coverage curves for a model trained on SST-2 and evaluated on each target. Selective-prediction performance, measured by AUC, deteriorates under shift. This is a property of the model’s raw confidence: monotone (positive) temperature scaling leaves AUC on a fixed dataset unchanged, so it is a failure separate from the calibration transfer gap. AUC also depends on the overall error rate, so we claim only that selective-prediction performance degrades, not that ranking quality alone deteriorated.

features, and fine-tuned transformers alike, and the target-fitted temperature stays low, showing that family inadequacy alone cannot explain the degradation: source-fitted recalibration parameters can fail to transfer when the target induces substantial confidence-space drift. But the cross-genre NLI result with RoBERTa (Table 2) shows the gap is not universal: when the shift is mild in the model’s confidence space, a source-fitted temperature shows no detectable average improvement from target refitting. The honest synthesis

Table 15 Keep-or-refit at target-label budget B : mean target-NLL improvement over keeping the source calibrator, with seed-level 95% bootstrap CIs, and the harmful-update rate at $B=25$ (fraction of pair-seed cases with NLL worse than keep by > 0.005). Design frozen on sentiment; the same rules are applied to ACS. “Always refit” is harmful at small B ; the evaluated rules reduce but do not eliminate the harm, and at $B=25$ none attains a positive mean on sentiment. The $\hat{C}$ criterion does not beat shrinkage/CV (see paired differences in text).

Task	Strategy	$B=25$	$B=50$	$B=100$	harm@25
Sentiment	always refit	-0.036 [-0.062, -0.016]	+0.003 [+0.000, +0.005]	+0.009 [+0.005, +0.012]	0.45
	shrink	-0.020 [-0.047, -0.003]	+0.004 [+0.003, +0.007]	+0.009 [+0.006, +0.012]	0.27
	CV	-0.021 [-0.046, -0.001]	-0.000 [-0.002, +0.002]	+0.006 [+0.003, +0.009]	0.20
	$\hat{C}$ rule	-0.030 [-0.055, -0.007]	-0.001 [-0.005, +0.002]	+0.007 [+0.003, +0.010]	0.30
ACS	always refit	-0.008 [-0.065, +0.024]	+0.031 [+0.021, +0.042]	+0.041 [+0.030, +0.049]	0.57
	shrink	+0.006 [-0.050, +0.039]	+0.036 [+0.028, +0.045]	+0.042 [+0.033, +0.050]	0.24
	CV	-0.000 [-0.076, +0.043]	+0.036 [+0.027, +0.044]	+0.041 [+0.032, +0.049]	0.14
	$\hat{C}$ rule	-0.002 [-0.061, +0.031]	+0.034 [+0.024, +0.044]	+0.041 [+0.031, +0.050]	0.22

Table 16 Decision-specific secondary metrics (mean over pair-seed cases) at $B=100$, confirming ECE and Brier move consistently with NLL. Keep is the source calibrator.

Task	Strategy	ECE	Brier	NLL
Sentiment	keep	0.053	0.178	0.535
	shrink	0.041	0.176	0.527
	$\hat{C}$ rule	0.044	0.177	0.528
ACS	keep	0.071	0.175	0.536
	shrink	0.047	0.164	0.494
	$\hat{C}$ rule	0.049	0.164	0.495

is conditional: in-domain calibration is an unreliable guarantee *when* the target induces a large shift in the model’s confidence behavior, and may remain transferable when that shift is small, as in our RoBERTa/MNLI setting. That this conditional pattern, and its parameter-drift association, reappears on an independent tabular benchmark with natural geographic shift (Section 6), where a single benchmark spans both a mild-shift/small-gap regime and a strong-shift/large-gap regime ordered by the parameter drift, provides evidence that it is a property of confidence-space drift under shift rather than of text data. This localizes both the problem and a direction: use unlabeled confidence divergence (Section 4) as an imperfect warning signal to prioritize target-label collection and recalibration when it becomes large, where the barrier is doing so without target labels. The external benchmark also shows this signal must be calibrated per modality: its rank-ordering of shifts transfers, but its absolute scale does not, so

a warning threshold tuned on one modality should not be applied verbatim to another.

Repairing it is non-trivial, for a structural reason, but not hopeless.

Propositions 1–2 explain why the natural label-free fixes lack a guarantee: mean confidence, accuracy, and variance do not identify the conditional-calibration curve, on which calibration error depends. Remark 1 sharpens the scope, the obstacle is recovering a map that is both calibrated *and* discriminative, not producing any calibrated output, and Proposition 3 supplies a sufficient condition for success: under correct specification mean-matching with an exact accuracy estimate does recover the temperature. The practical reading is therefore a boundary rather than a blanket impossibility, and two distinct ways the optimal fitted parameter can move are worth separating. Under *conditional-calibration drift*, the probability of correctness at a given score genuinely changes, so the target calibration map itself changes (this can occur while the map remains within the calibrator family or not); under *reweighting of an imperfect family*, the conditional correctness relationship is unchanged but different score regions become more common, so an imperfectly-fitting family’s optimal parameter moves even though the underlying curve did not. A moving optimal temperature therefore does not by itself prove the confidence–correctness relationship changed, and disentangling these two causes, e.g. with a small labelled target probe, is the most promising direction the diagnostic opens; researchers should

target label-free signals that constrain the *shape* of the relationship, not just its mean or spread.

Parameter drift is the strongest descriptive correlate, with important caveats; label-free signals are regime-dependent proxies.

A pattern across our settings is worth making explicit, framed as a two-level hierarchy. At the reference level, among the quantities we evaluate the optimal-temperature drift is the strongest descriptive correlate of the transfer gap in the shared-fit comparisons. Because the raw distance $|T_{\text{src}} - T_{\text{tgt}}|$ is *scale-dependent*, a common positive rescaling of the logits (and hence both temperatures) leaves calibrated probabilities unchanged yet changes it, we use the scale-invariant log-ratio $|\log(T_{\text{tgt}}/T_{\text{src}})|$ for cross-model and cross-task comparison. This invariant measure is associated with the gap at $\rho = 0.75$ on text and $\rho = 0.63$ on tabular data (both slightly stronger than the raw distance) and orders the mild and strong regimes both across tasks (sentiment vs. RoBERTa/MNLI) and within a single benchmark (mainland vs. Puerto-Rico shifts); no fitted temperature approached the optimizer bounds. Two caveats keep this claim honest. First, it is a *label-dependent reference diagnostic*, not a deployable indicator: computing T_{tgt} requires target labels, exactly the barrier Propositions 1–2 show mean-level statistics cannot circumvent. Second, these correlations are shared-fit (drift and gap use the same fitted temperatures), so they are partly mechanical: under independent estimation at 125 examples per fit the association falls to $\rho = 0.35$ and is not consistently positive across seeds (Section 4), with a no-shift reference showing temperature movement of comparable magnitude from sampling alone. We therefore read the drift as a descriptive *association* that strengthens with calibration sample size in the shared-fit setting but weakens and becomes seed-inconsistent under independent estimation, not a demonstrated causal mechanism or a validated predictor. At the deployable level, the label-free indicators are only proxies and are regime-dependent: the confidence-JS divergence, a deployable label-free proxy, transfers in rank but not absolute scale across modalities. (The ATC accuracy-estimation *error* is not a deployable signal at all, since computing it needs the

true target accuracy; we mention it only as a label-dependent diagnostic that happens to be weak on sentiment yet predictive on ACSIncome, underscoring that even label-dependent proxies are regime-dependent.) The reading is thus a qualified hierarchy: parameter drift is the strongest shared-fit correlate of transfer failure but a finite-sample-sensitive one, and confidence-JS is the best label-free but regime-dependent proxy that stands in for it when target labels are unavailable.

Implications for selective prediction and routing.

Systems that abstain, escalate, or route on a confidence threshold tuned in-domain will admit more errors under shift, as the threshold-transfer experiment shows directly. Coverage policies should be validated on target-like data or made adaptive, and the label-free divergence signal offers an imperfect way to help flag when revalidation is worthwhile.

Limitations.

We study binary sentiment and three-class NLI with models from linear classifiers to fully fine-tuned DistilBERT and RoBERTa, and three calibrator families (temperature, Platt, isotonic). The transfer gap is large and significant for sentiment-domain shift across this spectrum and across calibration metrics, and small for competent cross-genre NLI; we present this contrast as the finding rather than smoothing over it. Several extensions would sharpen the boundary. First, the NLI evaluation uses three seeds and models trained on moderate data on commodity hardware (mean in-domain accuracy 0.66, per-seed range 0.60–0.69); a higher-accuracy NLI model or harder genre pairs could probe whether the gap remains small as the model strengthens or the shift intensifies. Second, larger and generative models are untested; our claims are scoped to the encoder-classifier regime. Third, our controlled severity axis (Section 5) uses synthetic text corruptions; to corroborate the finding beyond NLP and beyond induced corruption, we replicate the gap and its parameter-drift association on an established tabular benchmark with natural geographic shift (Folktables ACSIncome, Section 6), where a single benchmark spans both regimes and the parameter drift orders them. Extending the controlled dose-response itself to a

graded image benchmark such as WILDS remains a useful further check. We consider natural domain shifts and controlled corruptions, not deliberately induced label shift or adversarial perturbation. Finally, our importance-weighted baselines are adaptations inspired by CPCS (Park et al. 2020), not faithful reproductions of its full framework: a confidence-only variant on the transformer and a stronger feature-based variant on the tabular benchmark (cross-fitted density ratios on the original ten features). Neither showed a detectable average ECE improvement, but this bounds these two variants rather than importance weighting in general, and a richer density-ratio estimator (e.g. over learned representations, or under an explicit covariate-shift assumption) could behave differently. The protocol, transfer-gap definition, and identifiability results are model- and task-agnostic.

9 Conclusion

We asked not *whether* calibration degrades under shift, which is known, but *why*, and answered by isolating the calibration transfer gap: the excess calibration error associated with transferring a source-fitted recalibration parameter relative to refitting within the same family on the target. Measuring the gap with strictly disjoint target-calibration and target-test splits, for sentiment-domain shift, across a model spectrum from linear classifiers to a fully fine-tuned transformer and three calibrator families, the gap is negligible in-domain and widens significantly under shift while a target-fitted temperature stays low, showing that parameter transfer contributes substantially to the observed degradation. Separately, and as a distinct failure, selective-prediction performance measured by AURC also deteriorates under shift; because positive temperature scaling preserves the binary confidence ordering, this failure cannot be addressed by positive temperature refitting and is not part of the transfer gap. For cross-genre NLI with a competent fine-tuned RoBERTa target refitting yields no detectable average improvement, showing the phenomenon is real but regime-dependent, closely associated with the drift in the optimal recalibration parameter. An independent tabular benchmark with natural geographic shift (Folktables ACSIncome), run through the identical leakage-free protocol, recreates both regimes

within one setting and reproduces the same parameter-drift ordering (scale-invariant $\rho = 0.63$, versus 0.75 on text), providing evidence that the gap and its association with parameter drift are not specific to NLP; the scale-invariant optimal-temperature drift is the strongest descriptive correlate of the gap among those evaluated when drift and gap share fitted parameters, but this association weakened under independent calibration estimation and varied across seeds, so we present it as a finite-sample-sensitive diagnostic rather than a validated predictor; the label-free proxies that stand in for it are likewise regime-dependent. We prove that mean confidence, variance, and accuracy do not identify the conditional-calibration curve, delimit this impossibility (a calibrated-but-uninformative constant map always exists) and give a sufficient condition for success (mean-matching recovers the temperature under correct family specification), and show empirically that three label-free heuristics, a confidence-only importance-weighted baseline (transformer), and a stronger feature-based one (tabular) all fail to beat the source calibrator, whereas 50–100 labelled target examples recover roughly half of the average 250-label-reference improvement in the linear sentiment setting. In-domain calibration is therefore an unreliable deployment guarantee *when* the confidence distribution shifts substantially. In the settings we tested, re-estimating the recalibration parameter from a modest amount of target supervision improved calibration, whereas the evaluated confidence-based label-free strategies did not; we do not claim that every successful approach must use target labels, only that the label-free strategies tested here were insufficient. On the operational question of whether to spend a small label budget refitting, our finding is partly negative: unconditional refitting on a few dozen labels is often harmful, uncertainty-aware rules (shrinkage, cross-validation, an information-criterion gate) reduce but do not eliminate that harm, and at the smallest budgets none reliably beats keeping the source calibrator, with the derived criterion not outperforming the simpler baselines. All code is provided as a supplementary archive for reproduction and extension.

Declarations

- **Funding.** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
- **Competing interests.** The author declares no competing interests.
- **Ethics approval and consent to participate.** Not applicable. This study uses only publicly available, de-identified benchmark and census microdata and involves no recruitment, intervention, or direct interaction with human participants.
- **Consent for publication.** Not applicable.
- **Data availability.** All datasets used are publicly available: SST-2 (GLUE), IMDB, Yelp Polarity, Rotten Tomatoes, and MultiNLI through the Hugging Face `datasets` library, and ACSIncome through the Folktables package (Ding et al. 2021), which retrieves US Census ACS PUMS data. No new data were collected.
- **Materials availability.** Not applicable.
- **Code availability.** All experiments are reproducible from the scripts described in Appendix A, provided as a supplementary code archive with this submission (environment specification, dataset versions, and saved per-seed split identifiers and results), which regenerate every reported number from the public datasets with fixed random seeds. A public repository will be made available on acceptance.
- **Author contributions.** B.S.N. conceived the study, designed and implemented the experiments, performed the analysis, and wrote the manuscript.

Appendix A Experimental details

Datasets and splits (four disjoint roles).

Sentiment domains are SST-2 (GLUE), IMDB, Yelp Polarity, and Rotten Tomatoes, all via the Hugging Face `datasets` library. To avoid any in-sample optimism in the target-fitted reference, each domain is divided into four disjoint roles: a *source-training* split (model fit), a *source-calibration* split (transferred calibrator), a *target-calibration* split (target-fitted reference calibrator), and an *untouched target-test* split (evaluation of both calibrators). The source-

and target-calibration splits are matched within each setting: $N_{\text{cal}} = 250$ for the linear, tabular, and fine-tuned-DistilBERT experiments, and $N_{\text{cal}} = 400$ for the three-class RoBERTa MNLI experiment. For the linear models the training pool is 4000 examples, of which 250 form the source-calibration split and the remaining 3750 fit the model; the target-calibration split is drawn from the target domain’s held-out pool (its training pool for the transformer/MNLI settings; the unused portion of the test pool for the linear settings), always disjoint from the target-test split. For the linear models the target-calibration and target-test splits are both carved, stratified and disjoint, from the same target evaluation pool, so a domain’s target-test size is (pool $- N_{\text{cal}}$) capped at $N_{\text{test}} = 1500$; no sampling with replacement is used. This matters for the two smaller corpora: SST-2’s public labelled split (GLUE validation, 872 examples) yields 250 target-calibration and 622 target-test examples, and Rotten Tomatoes’ test split (1066) yields 250 and 816; IMDB and Yelp have ample pools and reach the 1500 cap. (The larger transformer/MNLI settings instead draw target-calibration from the target’s disjoint *training* pool, leaving the full test split for evaluation; see below.) MNLI uses five genres (slate, telephone, government, fiction, travel) from `multi_nli`, with target-test drawn from the matched-validation split (test labels are not public) and the target-calibration split drawn from the (disjoint) training pool of the target genre. All headline gap, correlation, selective-prediction, and detection numbers are computed under this leakage-free protocol; the supplementary audit scripts (`run_transfer_audit.py`, `run_tabular_audit.py`, `run_finetune_audit.py`, `run_mnli_audit.py`, `run_frozen_audit.py`, `run_severity_audit.py`, `run_remedy_audit.py`, `run_labelfree_transformer_audit.py`) regenerate them, and a same-example target-fit reproduces the earlier, optimistic values for comparison. Per-seed split sizes are given in Table A1. For the transformer and MNLI settings the target-calibration split is drawn from the target domain’s training pool; because that pool is disjoint from both the model’s own training split and its source-calibration split (all three are formed by successive stratified partitions of the same pool

with fixed seeds), target-calibration examples are excluded from model training and from source-calibration by construction, and the released scripts assert this disjointness at run time.

External tabular benchmark.

The external replication (Section 6) uses ACSIncome from Folktabs (Ding et al. 2021) (2018 1-Year person survey), with US states as domains: the four largest states by ACS sample size (CA, TX, NY, FL) plus Puerto Rico (PR), included *a priori* as an expected strong demographic and socioeconomic shift. This domain set was fixed before any transfer-gap outcome was computed. Each state’s population is split 60%/40% into train and test pools; the four disjoint roles are formed as above, with a 4000-example training pool, of which 250 are reserved for source-calibration and 3750 are used to fit the model, and target-calibration ($N_{\text{cal}} = 250$) and an untouched target-test ($N_{\text{test}} \leq 1500$) carved from the test pool. The model is the ten standardized ACSIncome features with ℓ_2 logistic regression ($C = 1.0$), the linear analog of the sentiment LR model. All temperature/Platt/isotonic fitting, ECE/Brier/NLL, confidence-JS, $|T_{\text{src}} - T_{\text{tgt}}|$, ATC, and transfer-gap computations reuse the same functions as the sentiment experiment; only the data source and feature extractor differ. We use five seeds.

Controlled shift-severity corruption.

The severity experiment (Section 5) trains an LR classifier on clean IMDB and evaluates on the IMDB test split corrupted at severity $\delta \in \{0, \dots, 5\}$. Corruption is word-level on whitespace-split tokens. At severity δ a fraction 0.06δ of the tokens is perturbed (so 0, 6, 12, 18, 24, 30% for $\delta = 0, \dots, 5$); texts with fewer than four tokens are left unchanged. Each of the $n = \lceil 0.06\delta |\text{tokens}| \rceil$ perturbations independently and uniformly picks one of three operations at a uniformly random token position: (i) delete the token (only if more than four remain), (ii) swap it with the adjacent token, or (iii) substitute it with a token drawn uniformly from a fixed 20-word sentiment/function-word vocabulary. Stop-words and punctuation attached to tokens are eligible for corruption (no filtering). All corruption randomness is drawn from a NumPy generator

seeded by the experiment seed, so the corrupted target set is deterministic given (seed, δ). Following the leakage-free protocol, at each severity the corrupted target set is partitioned (stratified, fixed seed) into a target-calibration subset of $N_{\text{cal}} = 250$ examples, used only to fit T_{tgt} , and an untouched target-test subset on which the source- and target-fitted temperatures are both scored; the same partition is used across severities so the gap is measured on a consistent held-out test set. We use five seeds.

Models and hyperparameters.

LR: TF-IDF (max 20,000 features, uni+bigrams, sublinear tf, English stop words) with ℓ_2 logistic regression ($C = 4.0$, up to 2000 iterations). LSVM: the same TF-IDF with a linear SVM ($C = 1.0$) wrapped in Platt-style probability calibration (CalibratedClassifierCV, 3-fold). Frozen-feature model: DistilBERT [CLS] embeddings with a logistic-regression head. Fine-tuned sentiment model: **distilbert-base-uncased**, all layers, AdamW at learning rate 2×10^{-5} , batch size 16, 2 epochs, max sequence length 128. Fine-tuned NLI model: **roberta-base**, all layers, premise/hypothesis fed as a text pair, AdamW at learning rate 1×10^{-5} with weight decay 0.01, linear warmup over the first 10% of steps then linear decay, gradient-norm clipping at 1.0, batch size 16, 3 epochs, max sequence length 128, three seeds. (Warmup and clipping were necessary: without them RoBERTa fine-tuning intermittently collapsed to the three-class floor on some genres.)

Temperature and calibrator fitting.

Temperature T minimizes NLL on the relevant labels via bounded scalar optimization ($T \in [0.05, 10]$). Platt scaling fits an affine map $a z + b$ on the logit z by NLL. Isotonic regression is fit on the positive-class probability with out-of-bounds clipping. The target-fitted reference calibrator uses the same objective (NLL for temperature/Platt; the isotonic fit for isotonic) on target labels and is a diagnostic reference only, not a deployable method.

ATC and label-free strategies.

The ATC threshold t is chosen on the source calibration split so that the fraction of confidences above t equals source accuracy; the label-free

Table A1 Per-domain split sizes (per seed) under the four-role protocol. “Src-cal” and “Tgt-cal” are matched within each setting. Two split *sources* are used, as noted in the last column: linear/tabular settings carve target-calibration from the same evaluation pool as target-test (so target-test = pool − N_{cal} , capped at 1500), whereas the transformer/MNLI settings draw target-calibration from the target’s disjoint *training* pool, leaving the entire test split for evaluation. Src-train is the number of examples actually used to fit the model (the training pool minus the source-calibration carve-out).

Setting	Src-train	Src-cal	Tgt-cal	Tgt-test	Tgt-cal source
IMDB / Yelp (linear)	3750	250	250	1500	eval pool
SST-2 (linear)	3750	250	250	622	eval pool
Rotten Tomatoes (linear)	3750	250	250	816	eval pool
Fine-tuned DistilBERT	1250	250	250	up to 1000*	train pool
RoBERTa MNLI (per genre)	2100	400	400	1200	train pool
ACSIIncome (per state)	3750	250	250	≤1500	eval pool

*For fine-tuned DistilBERT, target-test is the subsampled test pool of size up to 1000; SST-2 as a target contributes its full 872-example validation split (no target-calibration is carved from the evaluation pool in this setting, since target-calibration comes from the training pool).

target-accuracy estimate is the fraction of target confidences above t . Mean-matching selects T so target mean confidence equals this estimate; variance-matching matches source confidence variance; pseudo-ECE minimizes a binned discrepancy against ATC-thresholded pseudo-correctness.

CPCS-style importance-weighted temperature.

Our importance-weighted baseline is *inspired by* the covariate-shift recalibration idea of Park et al. (2020): reweight the labeled source calibration set by an estimated density ratio $w(x) \approx p_t(x)/p_s(x)$ and fit temperature by weighted NLL on the reweighted source, so that the temperature is chosen for a source distribution reshaped toward the target. It is a lightweight adaptation, not a reproduction of the original: we estimate the density ratio with a logistic-regression domain classifier trained to separate source-calibration from (unlabeled) target points using the scalar max-confidence as the single feature, and convert its probability $\hat{p}(\text{target} \mid x)$ to weights $w = \hat{p}/(1 - \hat{p})$ clipped to $[10^{-3}, 10^3]$ and normalized to mean one. Temperature is then fit by minimizing the w -weighted NLL on the source calibration split (bounded scalar optimization, $T \in [0.05, 10]$). We use scalar confidence as the domain-classifier feature for this transformer comparison to maintain comparability with the other label-free signals. We do not claim this reproduces the original method’s full procedure or guarantees. *Feature-based variant* (Table 13). On the tabular benchmark we

additionally fit the domain classifier on the original ten ACSIIncome features (standardized by a scaler fit on the pooled source-calibration and target points; ℓ_2 -regularized logistic regression, $C = 1.0$). It separates the $N_{\text{cal}} = 250$ source-calibration points from the target-test points (≤ 1500), using their *features* but not their labels, a standard transductive covariate-shift design; the target-test labels remain reserved for evaluation, so this is a documented feature-access exception to the otherwise label- and data-free treatment of the target-test set. To avoid overfitting the weights to the source-calibration points they reweight, the density ratio is *cross-fitted* (two folds: each source point’s weight comes from a classifier trained on the other fold together with the target points), and weights $w = \hat{p}/(1 - \hat{p})$ are clipped a priori at the 99th percentile and normalized to mean one; the clip percentile and all hyperparameters are fixed without reference to target labels. Temperature is then fit by w -weighted NLL on the source-calibration split. The confidence-only ACS baseline in the same table uses a single-split (not cross-fitted) domain classifier on the scalar max-confidence with symmetric clipping to $[10^{-3}, 10^3]$, matching the transformer confidence-only baseline; we report the feature-based weights’ normalized effective sample size $\text{ESS}/n = (\sum_i w_i)^2 / (n \sum_i w_i^2)$ and maximum-to-mean ratio as concentration diagnostics.

Independent-estimation test (Table 6).

This uses the linear sentiment models (LR and LSVM) over the five primary seeds, on the shifted

source–target pairs. For each pair we take a *single* stratified partition of the source-calibration split and of the target-calibration split into halves A and B of 125 examples each (fixed by the seed; not repeated partitions), fit source and target temperatures separately on each half, estimate the drift $d_T^A = |\log(T_{\text{tgt},A}/T_{\text{src},A})|$ from half A , and compute the transfer gap from the independent half- B temperatures evaluated on the untouched target-test set. The same-group comparison uses half A for both drift and gap; the full-size reference uses the whole 250-example splits. The no-shift reference fits two independent source temperatures on halves A and B of the same source-calibration split and reports $|\log(T_A/T_B)|$.

Keep-or-refit decision and the bias–variance criterion (Section 7.4).

Write $\beta = 1/T$ and, for logit z and label $y \in \{0, 1\}$, the per-example NLL $\ell(\beta) = -y \log \sigma(\beta z) - (1 - y) \log(1 - \sigma(\beta z))$, with $\ell'(\beta) = z(\sigma(\beta z) - y)$ and $\ell''(\beta) = z^2 \sigma(\beta z)(1 - \sigma(\beta z)) \geq 0$, so $R(\beta) = \mathbb{E}[\ell]$ is convex. Let $\beta^* = \arg \min R$, $H = R''(\beta^*) = \mathbb{E}[z^2 \sigma(1 - \sigma)]$, and $J = \text{Var}(\ell'(\beta^*)) = \text{Var}(z(\sigma(\beta^* z) - y))$. The B -sample NLL minimiser $\hat{\beta}$ is the standard M -estimator. The expansion $\mathbb{E}[R(\hat{\beta}) - R(\beta^*)] \approx J/(2BH)$ and $\sqrt{B}(\hat{\beta} - \beta^*) \rightarrow \mathcal{N}(0, J/H^2)$ require: independent target observations; an *identifiable interior* population optimum β^* with *positive* curvature $H > 0$; and the smoothness and moment conditions for the standard expected-risk expansion. We stress these are genuine assumptions, not automatic consequences of bounded logits: a small calibration sample can favour a temperature at the optimizer boundary, and H can be near zero if the retained logits are small. Hence the expected out-of-sample improvement of refitting over the fixed source parameter is $\mathbb{E}[R(\beta_{\text{src}}) - R(\hat{\beta})] \approx D - J/(2BH)$ with $D = R(\beta_{\text{src}}) - R(\beta^*)$ (Eq. 5). The in-sample $\hat{D} = R_B(\beta_{\text{src}}) - R_B(\hat{\beta})$ overestimates D by $\approx J/(2BH)$, so the bias-corrected estimable criterion is $\hat{C} = \hat{D} - \hat{J}/(B\hat{H})$, with $\hat{H}, \hat{J}$ the empirical curvature and gradient variance at $\hat{\beta}$ on the B labels. This is a scalar information-criterion adjustment of the Takeuchi/AIC type (Konishi and Kitagawa 2008): in the correctly specified scalar case $J = H$ and the population correction reduces to $1/B$, the familiar one-parameter penalty. We use $\hat{C}$ heuristically as a keep-or-refit gate; it is an NLL-based

approximation that carries no finite-sample and no ECE guarantee, and we report ECE and Brier as separate empirical outcomes. *Numerical handling and diagnostics.* Temperatures are fit on $T \in [0.05, 10]$; across the 180 small-budget fits ($B \in \{25, 50, 100\}$), none reached the lower bound and 3 reached the upper bound, and the empirical curvature was rarely tiny (median $\hat{H} \approx 0.10$; below 10^{-3} in 0.6% of fits). The shrinkage weight $\lambda = \hat{D}^+ / (\hat{D}^+ + \hat{J}/(B\hat{H}))$ is defined as 0 when its denominator is non-positive (i.e. no refit), and $\hat{H}$ is floored at 10^{-8} to avoid division by zero. *Strategies.* “Shrink” uses $\lambda = \hat{D}^+ / (\hat{D}^+ + \hat{J}/(B\hat{H})) \in [0, 1]$ (with $\hat{D}^+ = \max(\hat{D}, 0)$), interpolating $\beta = \lambda \hat{\beta} + (1 - \lambda) \beta_{\text{src}}$; “CV” splits the B labels 70/30 (stratified when both classes have ≥ 4 examples, else it keeps the source), refits on the 70%, and picks keep-vs-refit by 30%-split NLL; the $\hat{C}$ rule refits iff $\hat{C} > 0$. The oracle uses the full 250-label target fit and is a reference, not a competitor. Design choices (the threshold 0 and the deterioration margin $\delta = 0.005$ used for harm/benefit rates) are fixed on the sentiment task and applied unchanged to the held-out ACS task; every label used for fitting, validation, or the decision counts toward B .

Threshold-transfer procedure.

For the selective-risk budget of 5%, the *source* threshold is the highest-coverage confidence value on the *source-calibration* split whose selective risk (error among retained points) does not exceed 5%; it is then applied unchanged to the untouched *target-test* split, where its realized risk and coverage are measured. The *oracle* threshold is selected the same way but on the disjoint *target-calibration* split (using its labels), and is likewise evaluated on the target-test split. Thus every threshold is selected on one split and evaluated on a strictly disjoint one; no threshold is chosen and scored on the same examples. Both realized risk and coverage on target-test are reported for each.

Metrics.

ECE uses equal-width bins at counts $\{10, 15, 20, 30, 50\}$; adaptive ECE uses equal-mass bins (15). We also report Brier score and NLL on the positive-class probability, and AURC (area under the risk–coverage curve).

Uncertainty quantification.

Primary linear-model experiments use five seeds {42, 1, 7, 123, 2024}; the fine-tuned DistilBERT (sentiment) and RoBERTa (NLI) experiments use three seeds each {42, 1, 7}. Confidence intervals are 95%. Effect-size CIs use a seed-level bootstrap (10,000 resamples): each quantity is first aggregated to one value per seed (averaging over the dependent source-target cells and models), then resampled over the independent seeds; these intervals describe variation across repeated runs on the fixed datasets and domains studied, not precision about arbitrary new domains. Spearman correlations over pooled shifted pairs are reported as descriptive summaries, accompanied by the range of the coefficient computed within each seed; we do not attach pooled-pair bootstrap confidence intervals to them, because the pooled pairs share source models and target sets and such intervals would understate that dependence.

Reproducibility.

All reported numbers are regenerated from a single entry point, `reproduce.py`, which runs the leakage-free audit scripts in order and writes each table’s source CSV; the mapping from tables/figures to scripts and output files is as follows. The core metric, calibrator, and shift-signal functions live in `run_calibration_transfer.py` and are imported unchanged by every audit script, so the protocol is identical across settings.

- Tables 1, 4, 5, 8 and Fig. 2 (sentiment linear): `run_transfer_audit.py` → `res_audit_*.csv`, `audit_analysis.txt`.
- Table 2 (RoBERTa MNLI): `run_mnli_audit.py` → `res_mnli_audit.csv`.
- Table 3 (model spectrum): `run_transfer_audit.py`, `run_frozen_audit.py`, `run_finetune_audit.py`.
- Table 10 (ACSIIncome): `run_tabular_audit.py` → `res_tabular_audit.csv`.
- Table 9 (severity): `run_severity_audit.py`; Table 11 and Table 14: `run_remedy_audit.py` and the budget block of `run_transfer_audit.py`; Table 12 (transformer label-free): `run_labelfree_transformer_audit.py`.

- Table 6 (independent-estimation test): `run_independent_drift.py`; Table 13 (feature-based weighting): `run_cpcs_features.py`; Table 15 and Fig. 1 (keep-or-refit decision): `run_adapt_decision.py`.

The earlier, pre-audit scripts (`run_calibration_transfer.py`, `run_severity.py`, `run_finetune.py`, `run_labelfree_transformer.py`, `run_mnli_transformer.py`, `run_external_tabular.py`) are retained only to reproduce the same-sample (optimistic) values quoted for contrast. A code archive with an environment specification, dataset versions, and saved split identifiers accompanies the submission; the repository URL / archive DOI will be included on acceptance and is available to reviewers on request.

References

- Alexandari A, Kundaje A, Shrikumar A (2020) Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In: International Conference on Machine Learning (ICML), pp 222–232
- Angelopoulos AN, Bates S, Candès EJ, et al (2021) Learn then test: Calibrating predictive algorithms to achieve risk control. In: arXiv preprint arXiv:2110.01052
- Brier GW (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review 78(1):1–3
- Clarté L, Loureiro B, Krzakala F, et al (2023) Expectation consistency for calibration of neural networks. In: Uncertainty in Artificial Intelligence (UAI)
- DeGroot MH, Fienberg SE (1983) The comparison and evaluation of forecasters. Journal of the Royal Statistical Society: Series D 32(1–2):12–22
- Ding F, Hardt M, Miller J, et al (2021) Retiring adult: New datasets for fair machine learning. In: Advances in Neural Information Processing Systems (NeurIPS), pp 6478–6490

- El-Yaniv R, Wiener Y (2010) On the foundations of noise-free selective classification. *Journal of Machine Learning Research* 11:1605–1641
- Garg S, Balakrishnan S, Lipton ZC, et al (2022) Leveraging unlabeled data to predict out-of-distribution performance. In: *International Conference on Learning Representations (ICLR)*
- Geifman Y, El-Yaniv R (2017) Selective classification for deep neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp 4878–4887
- Geifman Y, El-Yaniv R (2019) SelectiveNet: A deep neural network with an integrated reject option. In: *International Conference on Machine Learning (ICML)*, pp 2151–2159
- Gong Y, Lin X, Yao Y, et al (2021) Confidence calibration for domain generalization under covariate shift. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp 8958–8967
- Guillory D, Shankar V, Ebrahimi S, et al (2021) Predicting with confidence on unseen distributions. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp 1134–1144
- Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: *International Conference on Machine Learning (ICML)*, pp 1321–1330
- Hendrycks D, Dietterich T (2019) Benchmarking neural network robustness to common corruptions and perturbations. In: *International Conference on Learning Representations (ICLR)*
- Hendrycks D, Gimpel K (2017) A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: *International Conference on Learning Representations (ICLR)*
- Ji D, Logan RL, Smyth P, et al (2021) Active Bayesian assessment of black-box classifiers. In: *AAAI Conference on Artificial Intelligence*
- Jiang Y, Nagarajan V, Baek C, et al (2022) Assessing generalization of SGD via disagreement. In: *International Conference on Learning Representations (ICLR)*
- Kamath A, Jia R, Liang P (2020) Selective question answering under domain shift. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp 5684–5696
- Koh PW, Sagawa S, Marklund H, et al (2021) WILDS: A benchmark of in-the-wild distribution shifts. In: *International Conference on Machine Learning (ICML)*, pp 5637–5664
- Konishi S, Kitagawa G (2008) *Information Criteria and Statistical Modeling*. Springer
- Kull M, Perello-Nieto M, Kängsepp M, et al (2019) Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In: *Advances in Neural Information Processing Systems (NeurIPS)*
- Kumar A, Liang P, Ma T (2019) Verified uncertainty calibration. In: *Advances in Neural Information Processing Systems (NeurIPS)*
- Lipton ZC, Wang YX, Smola A (2018) Detecting and correcting for label shift with black box predictors. In: *International Conference on Machine Learning (ICML)*, pp 3122–3130
- Liu Y, Ott M, Goyal N, et al (2019) RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*
- Minderer M, Djolonga J, Romijnders R, et al (2021) Revisiting the calibration of modern neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS)*
- Nguyen K, O’Connor B (2015) Posterior calibration and exploratory analysis for natural language processing models. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp 1587–1598
- Nixon J, Dusenberry MW, Zhang L, et al (2019) Measuring calibration in deep learning. In: *CVPR Workshops*

- Ovadia Y, Fertig E, Ren J, et al (2019) Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems (NeurIPS)
- Pampari A, Ermon S (2020) Unsupervised calibration under covariate shift. arXiv preprint arXiv:200616405
- Park S, Bastani O, Weimer J, et al (2020) Calibrated prediction with covariate shift via unsupervised domain adaptation. In: International Conference on Artificial Intelligence and Statistics (AISTATS), pp 3219–3229
- Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers 10(3):61–74
- Quiñonero-Candela J, Sugiyama M, Schwaighofer A, et al (2009) Dataset Shift in Machine Learning. MIT Press
- Roelofs R, Cain N, Shlens J, et al (2022) Mitigating bias in calibration error estimation. In: International Conference on Artificial Intelligence and Statistics (AISTATS), pp 4036–4054
- Sanh V, Debut L, Chaumond J, et al (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:191001108
- Wang X, Long M, Wang J, et al (2020) Transferable calibration with lower bias and variance in domain adaptation. In: Advances in Neural Information Processing Systems (NeurIPS)
- Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp 694–699
- Zhang J, Kailkhura B, Han TYJ (2020) Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In: International Conference on Machine Learning (ICML), pp 11117–11128

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [jdsasupplementarycode.zip](#)