CALIBRATING FINE-TUNED MODELS VIA SELECTIVE CONFIDENCE REUSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning can improve a model's predictions without making its confidence a reliable guide to whether those predictions are correct. Calibration based only on the final confidence score cannot distinguish equally confident predictions with different probabilities of being correct, even when the model before fine-tuning retains information that separates them. We recast calibration as selective inheritance: preserve reliable information from the model before fine-tuning (the predecessor) where confidence remains stable, and repair confidence estimates elsewhere. Our method, A-PAIR, aligns predecessor scores into a reference and uses labeled source data and unlabeled target inputs to audit where this reference can be retained and where correction is needed. Our theory shows that predecessor information can recover distinctions inaccessible to score-only calibration, and that selective inheritance can reduce finite-sample estimation error when repair is localized. On public benchmarks under controlled distribution shifts, A-PAIR improves the calibration of fine-tuned language models while keeping their predictions accurate.

1 INTRODUCTION

Fine-tuning has become the default way to adapt pretrained models to downstream tasks (Devlin et al., 2019; Bommasani et al., 2021). A model is first (pre-)trained on large and diverse datasets, then updated by supervised fine-tuning, instruction tuning, or another post-training procedure before deployment (Wei et al., 2022; Ouyang et al., 2022). We refer to the model before this update as the *predecessor model* and to the updated model as the *fine-tuned model*. High predictive accuracy, however, does not ensure that the model's confidence accurately reflects the probability of a correct answer. In particular, fine-tuning may distort the confidence on some parts of the input space, even when the predecessor is well calibrated (OpenAI, 2023; Kumar et al., 2022; Kong et al., 2020; He et al., 2023; Zhu et al., 2023). We study confidence calibration for the event that the answer selected by the fine-tuned model is correct, keeping that answer fixed and computing the predecessor's confidence for the same answer. In the setting considered here, both models are retained and can be evaluated on the same inputs. Calibration after fine-tuning then asks how information from both models can be used to estimate the probability that the selected answer is correct, without changing the answer itself.

Score-only calibration methods, including temperature scaling, adjust predicted probabilities using the information retained in the final fine-tuned score (Guo et al., 2017). Fine-tuning, however, may collapse distinctions that remain visible in the predecessor. For example, suppose the selected answer is correct with probability 0.2 for one group of inputs and 0.8 for another, but fine-tuning gives both groups the same logit. Any rule that sees only that logit must assign the same probability to both groups, however flexible its correction function. This probability can be calibrated for the pooled group without expressing the difference in correctness probabilities between its subgroups. The limitation lies in the information available from the final score, rather than the expressive power of the correction. If the predecessor logit still separates the groups, the score pair can retain additional label-relevant information, as formalized in Prop. 2.1. Using the predecessor score may allow probability predictions to reflect differences that the final score alone cannot represent.

Retaining label-relevant distinctions does not ensure that the predecessor assigns correct probabilities. Yet where a predecessor-derived *anchor* is reliable, re-estimating probabilities from scratch

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

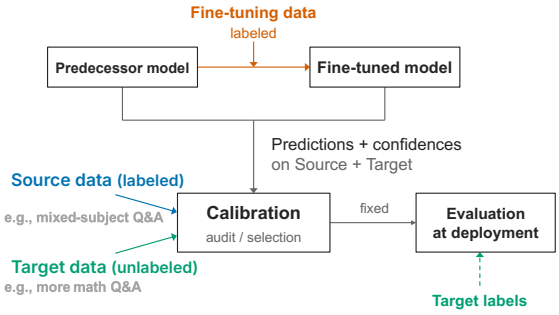


Figure 1: Post-fine-tuning calibration setting

may increase finite-sample error. Thus, we seek to keep predictions near the anchor where supported by an audit and implement label-based repair elsewhere. We study this principle in a deployment setting without target labels: after both models are fixed, held-out labeled source calibration data and unlabeled target inputs are available. The source calibration data serve a different role from the fine-tuning data and need not follow the same distribution. Both models can be evaluated on target inputs, but target labels are excluded from calibrator construction, auditing, and selection. Such asymmetric label access arises in established calibration settings (Park et al., 2020; Wang et al., 2020), including fraud detection (Carcillo et al., 2018), clinical settings (Campanella et al., 2025), and recent LLM calibration (Luo et al., 2025). Because source and target input distributions may differ, we weight source data to account for the target input distribution; Sec. 5.3 states the conditions under which this yields target validity. The question is thus how to audit the anchor with the available data, reuse it where supported, and repair the remaining regions using source labels.

We formalize post-fine-tuning calibration as an inheritance-or-repair problem and develop A-PAIR (audited predecessor-anchored inheritance-or-repair). We align the predecessor score to the fine-tuned scale to obtain an *anchor*, then use the discrepancy between the fine-tuned logit and the anchor to propose *stable* regions for preservation and *drift* regions for subgroup-level repair. We use source labels to check whether each candidate passes the anchor audit, without assuming that either model is uniformly reliable. The key is to separate auditing the anchor with source labels from measuring departures from it on unlabeled target inputs. On the drift region, updates follow the Brier-descent procedure behind multicalibration and regression boosting (Globus-Harris et al., 2023). Our contributions and paper structure are summarized as follows.

- **A new view of calibration after fine-tuning.** We show that the score pair can retain label-relevant distinctions that no correction of the final score alone can recover, and cast post-fine-tuning calibration as an inheritance-or-repair problem (Secs. 2 and 3).
- **A selective inheritance-or-repair method.** We turn this insight into a deployable procedure that reuses predecessor information, repairs the rest using labeled source data, and falls back to repair when preservation is not supported (Sec. 4).
- **Statistical benefit and auditable guarantees.** We show that selective inheritance can reduce finite-sample re-estimation cost when only a low-complexity, initially unknown region requires repair. Separately, we give a finite-sample selection guarantee for an observable calibration criterion relative to eligible candidates on the selected path, and state the source-to-target conditions under which this guarantee controls audited target calibration error (Sec. 5).
- **Evidence in language-model settings.** Under subject-mixture shifts on MMLU-Pro, A-PAIR uses both inheritance and repair and achieves the lowest calibration error among the compared methods, without target labels for construction or selection. On SciQ, where simple score-only recalibration already works well, the simpler baseline achieves lower calibration error (Sec. 6).

The appendix contains related literature, proofs, and further implementation details.

2 Why the Fine-Tuned Score Is Not Enough

Fine-tuning can coarsen the information available to score-only post-processing: distinct label-relevant regions may collapse to the same fine-tuned score. This section first uses Brier risk (Murphy, 1973; Hébert-Johnson et al., 2018) to formalize the resulting loss of resolution.

Consider the binary setting induced by the answer $\hat{j}$ selected by the fine-tuned model, with $Y = \mathbf{1}\{\hat{j} \text{ is correct}\}$. Let $q_{\mathrm{pre}}(x)$ and $q_{\mathrm{ft}}(x)$ denote the probabilities assigned to this same selected option $\hat{j}$ by the predecessor and fine-tuned models, respectively, and define $z_{\mathrm{pre}}(x) := \mathrm{logit}(q_{\mathrm{pre}}(x))$

and $z_{\mathrm{ft}}(x) := \mathrm{logit}(q_{\mathrm{ft}}(x))$. Let P_{tgt} be the target distribution over $(X, Y) \in \mathcal{X} \times \{0, 1\}$, and let $\eta(x) := P_{\mathrm{tgt}}(Y = 1 \mid X = x)$ be the target conditional probability. A *score-only or blind* correction is any post-processing rule that only sees the fine-tuned logit, so it has the form $T(z_{\mathrm{ft}}(x))$ for some function $T : \mathbb{R} \to [0, 1]$. A *pair-aware* correction may use both logits and has the form $R(z_{\mathrm{ft}}(x), z_{\mathrm{pre}}(x))$ for some function $R : \mathbb{R} \times \mathbb{R} \to [0, 1]$. We compare them using the Brier risk $\mathsf{Risk}_P(q) := \mathbb{E}_P[(Y - q(X))^2]$.

Brier risk provides a clear way to compare these two classes of correction rules. For the information S available to the post-processing rule, let $\eta_S := \mathbb{E}[Y \mid S]$ and $\bar{\eta} := \mathbb{E}[Y]$. Any predictor $q(S)$ satisfies the following decomposition (Brier, 1950; Murphy, 1973):

$$\mathsf{Risk}_P(q(S)) = \underbrace{\bar{\eta}(1 - \bar{\eta})}_{\text{uncertainty}} - \underbrace{\mathbb{E}[(\eta_S - \bar{\eta})^2]}_{\text{resolution}} + \underbrace{\mathbb{E}[(q(S) - \eta_S)^2]}_{\text{reliability}}.$$

Thus recalibration can reduce the reliability term relative to the information S, but it cannot increase the resolution of that state. Our claim is that after fine-tuning, the fine-tuned logit $S_{\mathrm{ft}} = z_{\mathrm{ft}}(X)$ can be a coarsened state, while the pair $S_{\mathrm{pair}} = (z_{\mathrm{ft}}(X), z_{\mathrm{pre}}(X))$ may still preserve label-relevant resolution.

Proposition 2.1 (Score-only correction cannot recover resolution lost during fine-tuning)**.** *Let $C_0, C_1 \subseteq \mathcal{X}$ be disjoint measurable sets, with $p_i := P_{\mathrm{tgt}}(C_i) > 0$ and $\bar{\eta}_i := \mathbb{E}[Y \mid X \in C_i]$. Assume that $z_{\mathrm{ft}} = v$ on $C_0 \cup C_1$ for some $v \in \mathbb{R}$, that $\bar{\eta}_0 \neq \bar{\eta}_1$, and that there exist disjoint Borel sets $J_0, J_1 \subseteq \mathbb{R}$ such that $C_i = z_{\mathrm{pre}}^{-1}(J_i)$ for $i = 0, 1$. Then, for every $T : \mathbb{R} \to [0, 1]$,*

$$\mathsf{Risk}_{P_{\mathrm{tgt}}}(T(z_{\mathrm{ft}}(X))) - \inf_{R:\mathbb{R}^2 \to [0,1]} \mathsf{Risk}_{P_{\mathrm{tgt}}}(R(z_{\mathrm{ft}}(X), z_{\mathrm{pre}}(X))) \geq \frac{p_0 p_1}{p_0 + p_1}(\bar{\eta}_0 - \bar{\eta}_1)^2.$$

Interpretation. On $C_0 \cup C_1$, the fine-tuned score permits only one prediction, so every score-only post-processing rule incurs an excess Brier risk of at least the displayed resolution term. If the predecessor still separates the two subpopulations C_0 and C_1, the score pair retains a label-relevant distinction that no post-processing of the fine-tuned score alone can recover.

From retained resolution to selective inheritance. Prop. 2.1 isolates an information limit under exact score collapse; the diagnostics on a language model in Sec. 6 provide empirical motivation at similar fine-tuned confidence. An informational advantage does not ensure finite-sample gains: refitting a flexible pair-aware calibrator everywhere may still require relearning the full probability model. We instead align the predecessor into an anchor, inherit it where supported by audits, and restrict label-based repair elsewhere. Sec. 5.1 analyzes when this procedure helps.

3 Anchors and the Stable/Drift Split

To make selective inheritance operational, this section defines the anchor and the stable/drift split.

Let P_{src} be the labeled source calibration distribution. $\hat{w} : \mathcal{X} \to [0, W]$ denotes a clipped estimate of the target-over-source density ratio $dP_{\mathrm{tgt}}^X / dP_{\mathrm{src}}^X$, where W is its clipping level.

Anchor. The predecessor logit may retain distinctions lost after fine-tuning, but its numerical scale need not match that of the fine-tuned logit. Hence, we fit an increasing map $\hat{\psi} : \mathbb{R} \to \mathbb{R}$ (e.g., by least squares) to align the predecessor logits to the fine-tuned logit scale on unlabeled target construction inputs (App. F.1), and define

$$\text{anchor logit: } z^{\mathsf{a}}(x) := \hat{\psi}(z_{\mathrm{pre}}(x)), \qquad \text{anchor probability: } q^{\mathsf{a}}(x) := \mathrm{sigmoid}(z^{\mathsf{a}}(x)).$$

The anchor q^{a} is a scale-aligned reference that preserves the predecessor's ordering information. It provides a common logit scale for measuring the discrepancy between the predecessor and fine-tuned model and, subsequently, for deciding where preservation should be considered. Note that the anchor is a reference and we do not assume that it is calibrated.

Stable/drift split. For a threshold $\tau > 0$, define a stable-region indicator

$$s_\tau(x) := \mathbf{1}\{|z_{\mathrm{ft}}(x) - z^{\mathsf{a}}(x)| \leq \tau\}.$$

It defines an observable stable/drift split: points with $s_\tau(x) = 1$ lie in the stable region, while those with $s_\tau(x) = 0$ lie in the drift region. The discrepancy bounds the cost of preservation: since the sigmoid is $1/4$-Lipschitz, $s_\tau(x) = 1$ implies $|q_{\mathrm{ft}}(x) - q^{\mathsf{a}}(x)| \leq \tau/4$. A larger τ increases stable-region coverage, but requires source-to-target transfer to hold over a larger region. Thus, stability only defines the partition between the preservation and repair regions. Calibration on the target distribution additionally requires the transfer conditions (see Sec. 5.3).

4 METHOD

Overview. A-PAIR proceeds in four steps (Alg. 1). First, we prepare candidate anchors, stable regions, and finite classes of audit functions. Second, we construct a predictor path for each candidate θ through preservation on stable regions and repair on drift regions. Third, we use independent holdouts to audit the anchors and select the candidate with the largest target coverage among those that pass the anchor audit. Fourth, we apply a source-Brier screen and minimize $\widehat{C}_{\mathrm{obs}}$ within the selected path. A path is a finite sequence of saved predictors, not just its final iterate. Its construction follows the Brier-descent template underlying multicalibration and boosting (Hébert-Johnson et al., 2018; Globus-Harris et al., 2023; Gopalan et al., 2022a).

Data protocol. Let S_{const} and S_{ho} be the labeled source construction and holdout samples. Write U_{const} and U_{ho} for their unlabeled target counterparts. Preprocessing and path construction use only construction data. All candidate predictors, weights, thresholds, bin boundaries, and resulting audit functions are fixed before either independent holdout is inspected. Source holdout labels support anchor, repair, and Brier audits; target holdout covariates support coverage and preservation checks. Holdouts are used only for auditing and selection; target labels are reserved for final evaluation (see Tab. 2 for the data splits and their roles in construction, auditing, and selection).

Step 1: Define candidate anchors, stable regions, and audit functions. Let $\theta \in \Theta$ index a finite collection of candidate anchors and stable regions. Each candidate θ determines an anchor $q_{\mathrm{anc},\theta}$ and a stable/drift partition through s_θ. We include the all-repair candidate θ_{rep} with $s_{\theta_{\mathrm{rep}}} \equiv 0$. We also fix $\hat{w} : \mathcal{X} \to [0, W]$, positive construction thresholds $\gamma_{\mathrm{pres}}, \gamma_{\mathrm{rep}}$, anchor tolerance γ_{anc}, and iteration cap T.

For each $\theta \in \Theta$, we specify finite classes of audit functions $\mathcal{A}_{\mathrm{pres}}(\theta)$ and $\mathcal{A}_{\mathrm{rep}}(q, \theta)$. Suppressing θ, we write $q^{\mathsf{a}} = q_{\mathrm{anc},\theta}$ and $s = s_\theta$. Representative stable- and drift-region tests are

$$a_{g,\phi}(x) = s(x)g(x)\phi(q^{\mathsf{a}}(x)), \qquad b_{q,g,I}(x) = (1 - s(x))g(x)\mathbf{1}\{q(x) \in I\}.$$

Here g is a prespecified subgroup test that may use the paired model scores, ϕ localizes tests along the anchor-probability axis, and I is a prespecified repair-score interval. We normalize g and ϕ so that $\|g\|_\infty, \|\phi\|_\infty \le 1$, ensuring that all tests are bounded by one. The preservation class is fixed by the anchor and stable region; the repair class changes with q through score-bin membership. These same classes determine construction updates and subsequent holdout audits. For θ_{rep}, we omit preservation and use only repair tests.

Step 2: Construct a predictor path for each candidate θ. For each $\theta \in \Theta \cup \{\theta_{\mathrm{rep}}\}$, we initialize $q_\theta^{(0)}$ (e.g., $q_\theta^{(0)} \equiv 0$) and construct $\mathcal{Q}_\theta$ using only construction data. Preservation reduces deviations from the anchor on the stable region, while repair reduces label-based residuals on the drift region. The construction potential captures these two roles as follows:

$$\widehat{L}(q) := \underbrace{\hat{\mathbb{E}}_{U_{\mathrm{const}}}[s(X)\{q(X) - q^{\mathsf{a}}(X)\}^2]}_{=:\widehat{L}_{\mathrm{pres}}(q)} + \underbrace{\hat{\mathbb{E}}_{S_{\mathrm{const}}}[\hat{w}(X)(1 - s(X))\{Y - q(X)\}^2]}_{=:\widehat{L}_{\mathrm{rep}}(q)}.$$

The first term encourages agreement with the anchor using unlabeled target inputs; the second is the density-ratio-weighted source Brier loss on the drift region. We define the population violations

$$V_{\mathrm{pres}}(q) := \sup_{a \in \mathcal{A}_{\mathrm{pres}}} \left|\mathbb{E}_{P_{\mathrm{tgt}}}[a(X)\{q(X) - q^{\mathsf{a}}(X)\}]\right|,$$
$$V_{\mathrm{rep}}(q) := \sup_{b \in \mathcal{A}_{\mathrm{rep}}(q)} \left|\mathbb{E}_{P_{\mathrm{src}}}[\hat{w}(X)b(X)\{Y - q(X)\}]\right|.$$

Here, $V_{\mathrm{pres}}(q)$ measures target-side departure from the anchor, while $V_{\mathrm{rep}}(q)$ measures weighted source residuals on the drift region.

Path construction uses empirical versions of these violations on the construction samples, never on the holdouts. At round t, we choose $a_t \in \mathcal{A}_{\mathrm{pres}}$ and $b_t \in \mathcal{A}_{\mathrm{rep}}(q^{(t)})$ that maximize the absolute values of the corresponding empirical correlations on the construction samples. Their signed correlations are $c_t^{\mathrm{pres}} := \hat{\mathbb{E}}_{U_{\mathrm{const}}}[a_t(X)\{q^{(t)}(X) - q^{\mathsf{a}}(X)\}]$ and $c_t^{\mathrm{rep}} := \hat{\mathbb{E}}_{S_{\mathrm{const}}}[\hat{w}(X)b_t(X)\{Y - q^{(t)}(X)\}]$. The preservation and repair updates are, respectively,

$$q_{\mathrm{pres}}^{(t+1)} := \mathrm{proj}_{[0,1]}\left(q^{(t)} - c_t^{\mathrm{pres}} a_t\right), \qquad q_{\mathrm{rep}}^{(t+1)} := \mathrm{proj}_{[0,1]}\left(q^{(t)} + \frac{c_t^{\mathrm{rep}}}{W} b_t\right). \tag{4.1}$$

Algorithm 1 A-PAIR

Require: Labeled source $S_{\text{const}}, S_{\text{ho}}$; unlabeled target $U_{\text{const}}, U_{\text{ho}}$, with independent holdouts; paired model logits.

Require: Preprocessing rules for anchors, stable regions, audit functions, and $\hat{w} : \mathcal{X} \to [0, W]$; construction thresholds $\gamma_{\text{pres}}, \gamma_{\text{rep}} > 0$; anchor tolerance γ_{anc}; cap T; simultaneous audit-bound rules at confidence level $1 - \delta$, $\delta \in (0, 1)$.

Ensure: Selected candidate $\hat{\theta}$ and confidence predictor $\hat{q} : \mathcal{X} \to [0, 1]$.

Step 1: Define candidate anchors, stable regions, and audit functions.

Using construction data, fix $\hat{w}$, candidate set $\Theta \cup \{\theta_{\text{rep}}\}$, anchors, stable-region indicators, and rules for constructing the audit functions.

Step 2: Construct a predictor path for each candidate θ.

for $\theta \in \Theta \cup \{\theta_{\text{rep}}\}$ **do**
  Initialize $q_\theta^{(0)}$ and set $\mathcal{Q}_\theta \leftarrow \{q_\theta^{(0)}\}$.
  for $t = 0, \ldots, T-1$ **do**
    Find maximizing tests and signed construction correlations $c_t^{\text{pres}}, c_t^{\text{rep}}$ as in Step 2; set $c_t^{\text{pres}} = 0$ for θ_{rep}.
    if $|c_t^{\text{pres}}| \le \gamma_{\text{pres}}$ and $|c_t^{\text{rep}}| \le \gamma_{\text{rep}}$ **then**
      break
    end if
    Among threshold-exceeding updates, choose the larger decrease guarantee, $(c_t^{\text{pres}})^2$ or $(c_t^{\text{rep}})^2/W$, and apply Eq. (4.1).
    Append $q_\theta^{(t+1)}$ to $\mathcal{Q}_\theta$.
  end for
end for

Step 3: Audit the anchors and select a candidate.

Compute simultaneous audit radii for the full fixed candidate family (Thm. 5.2; App. D.1).

On S_{ho}, retain candidates satisfying $\widehat{V}_{\text{anc}}(\theta) + \xi_{\text{anc}} \le \gamma_{\text{anc}}$.

Among the candidates that pass the anchor audit, select the one with the largest coverage on U_{ho} by Eq. (4.3); use θ_{rep} if none passes.

Step 4: Screen and select the final predictor.

Set $\mathcal{E}_{\hat{\theta}} \leftarrow \{q_{\hat{\theta}}^{(0)}\} \cup \{q \in \mathcal{Q}_{\hat{\theta}} : U_{\text{src}}(q; \hat{\theta}) \le 0\}$ using S_{ho}.

Set $\hat{q} \in \arg\min_{q \in \mathcal{E}_{\hat{\theta}}} \widehat{C}_{\text{obs}}(q)$ using Eq. (4.4) and the repair-only convention for θ_{rep}.

return $(\hat{\theta}, \hat{q})$.

Preservation is eligible when $|c_t^{\text{pres}}| > \gamma_{\text{pres}}$; repair is eligible when $|c_t^{\text{rep}}| > \gamma_{\text{rep}}$. We apply at most one eligible update, choosing the larger guaranteed decrease in $\widehat{L}$: $(c_t^{\text{pres}})^2$ for preservation or $(c_t^{\text{rep}})^2/W$ for repair. We stop when neither is eligible or T updates have been applied, retaining the initializer and every iterate in $\mathcal{Q}_\theta$. Positive thresholds and $0 \le \widehat{L} \le 1 + W$ imply finite termination even without the cap (App. D.1). After constructing $\mathcal{Q}_\theta$, we instantiate and freeze $\mathcal{A}_{\text{rep}}(q, \theta)$, including all current-score bin memberships, for every saved $q \in \mathcal{Q}_\theta$.

Step 3: Audit the anchors and select a candidate. After all paths are fixed, we evaluate the candidate anchors indexed by $\theta \in \Theta$ on S_{ho}:

$$\widehat{V}_{\text{anc}}(\theta) := \sup_{a \in \mathcal{A}_{\text{pres}}(\theta)} \left| \widehat{\mathbb{E}}_{S_{\text{ho}}} \left[\hat{w}(X) a(X) \{Y - q_{\text{anc},\theta}(X)\} \right] \right|. \tag{4.2}$$

Replacing the sample average in Eq. (4.2) with $\mathbb{E}_{P_{\text{src}}}$ defines $V_{\text{anc}}(\theta)$, the largest weighted source residual of the anchor. A candidate θ passes the anchor audit if $\widehat{V}_{\text{anc}}(\theta) + \xi_{\text{anc}} \le \gamma_{\text{anc}}$, where ξ_{anc} is a simultaneous audit radius. Among the candidates that pass the anchor audit, A-PAIR selects

$$\hat{\theta} \in \underset{\theta \in \Theta:\, \widehat{V}_{\text{anc}}(\theta) + \xi_{\text{anc}} \le \gamma_{\text{anc}}}{\arg\max} \widehat{P}_{\text{tgt}}\{s_\theta(X) = 1\}, \tag{4.3}$$

with coverage measured on U_{ho}. If none passes, we set $\hat{\theta} := \theta_{\text{rep}}$. Among the passing candidates, larger coverage assigns more target mass to inheritance rather than label-based re-estimation. Thus, we use source labels to check whether each candidate passes the anchor audit and target covariates

to determine coverage. Stable-region selection prioritizes reuse subject to the anchor audit; the observable criterion is minimized separately within the selected path $\mathcal{Q}_{\hat{\theta}}$.

Step 4: Screen and select the final predictor. Path construction reduces the preservation and repair potential, not overall predictive risk. For $q \in \mathcal{Q}_\theta$, let $\widehat{\Delta}_{\mathrm{src}}(q;\theta) := \hat{\mathbb{E}}_{S_{\mathrm{ho}}}[(Y-q(X))^2 - (Y-q_\theta^{(0)}(X))^2]$ be its paired source-Brier difference from the initializer. We form the simultaneous one-sided upper bound $U_{\mathrm{src}}(q;\theta) := \widehat{\Delta}_{\mathrm{src}}(q;\theta) + \xi_{\mathrm{src}}$, with source-Brier radius ξ_{src} as in App. D.1. Let $\mathcal{E}_\theta$ retain the initializer unconditionally and other candidates only when $U_{\mathrm{src}}(q;\theta) \le 0$, ruling out an increase in source Brier risk. We apply this screen to the selected path $\mathcal{Q}_{\hat{\theta}}$. It does not certify the initializer itself or guarantee target Brier risk.

For the selected candidate $\hat{\theta}$, the stable-region identity $Y - q = (Y - q^{\mathsf{a}}) + (q^{\mathsf{a}} - q)$ separates the label-dependent anchor residual from the departure measurable on unlabeled target inputs. The triangle inequality motivates the preservation-side sum; taking the worse of the two sides gives

$$C_{\mathrm{obs}}(q) := \max\{V_{\mathrm{anc}} + V_{\mathrm{pres}}(q),\ V_{\mathrm{rep}}(q)\}.$$

We evaluate its uncertainty-adjusted counterpart

$$\widehat{C}_{\mathrm{obs}}(q) := \max\left\{\widehat{V}_{\mathrm{anc}} + \xi_{\mathrm{anc}} + \widehat{V}_{\mathrm{pres}}(q) + \xi_{\mathrm{pres}},\ \widehat{V}_{\mathrm{rep}}(q) + \xi_{\mathrm{rep}}\right\}, \tag{4.4}$$

where anchor and repair terms use S_{ho}, preservation uses U_{ho}, and the ξ terms are simultaneous radii from Thm. 5.2. For all-repair, we use $C_{\mathrm{obs}}(q) = V_{\mathrm{rep}}(q)$ and $\widehat{C}_{\mathrm{obs}}(q) = \widehat{V}_{\mathrm{rep}}(q) + \xi_{\mathrm{rep}}$. A-PAIR returns $\hat{q} \in \arg\min_{q \in \mathcal{E}_{\hat{\theta}}} \widehat{C}_{\mathrm{obs}}(q)$, which need not be the last iterate. The comparison in Thm. 5.2 is within this selected path, not across candidates θ.

5 Theoretical Results

We study three questions: when selective inheritance can reduce finite-sample learning cost in a finite-state model (Sec. 5.1), how A-PAIR's holdout selection is controlled (Sec. 5.2), and when observable audits imply target calibration (Sec. 5.3).

5.1 When does the preserve-or-repair structure help?

Selective inheritance can avoid relearning probabilities that the anchor already estimates correctly. To isolate this benefit, we analyze an auxiliary audited preserve-or-repair selector that retains A-PAIR's repair updates but uses a specialized stopping rule and quadratic certificate for joint region and predictor selection (App. E.5).

Theorem 5.1 (Statistical advantage of selective inheritance; informal version of Thm. E.11)**.** *Consider a label-independent, equal-mass K-cell discretization of the observed score pair, with source and target coinciding. Let $X \sim \mathrm{Unif}([K])$, $\eta_j = \mathbb{P}(Y=1 \mid X=j)$, and let $a \in [1/4, 3/4]^K$ be a fixed, label-independent anchor. Suppose that $\eta \in [1/4,3/4]^K$ and that there exists an unknown repair region D in a fixed finite candidate family $\mathcal{D}$, with $|D| \le k_{\mathrm{rep}}$, such that $\eta_j = a_j$ for every $j \notin D$. With n labeled observations, the audited preserve-or-repair selector achieves*

$$\mathbb{E}[\mathcal{E}_\eta(\widehat{q}_{\mathrm{str}})] \lesssim \underbrace{\frac{k_{\mathrm{rep}}}{n}}_{\text{repair estimation}} + \underbrace{\frac{\sqrt{K}\,\ell}{n}}_{\text{region selection and audit}},$$

where $\mathcal{E}_\eta(q) = K^{-1}\sum_{j=1}^K (q_j - \eta_j)^2$ is excess Brier loss and ℓ is logarithmic in the number of resulting candidates, n, and $1/\delta$. In contrast, independently refitting all K cells with the same label budget incurs $\mathbb{E}[\mathcal{E}_\eta(\widehat{q}_{\mathrm{full}})] \gtrsim \frac{K}{n}$.

The key proof step is a block-median order-two U-statistic that removes diagonal bias and enables a single uniform comparison across candidate partitions and their path predictors (see App. E).

Interpretation. The theorem quantifies the benefit of reducing the effective estimation problem from all K cells to the cells that actually require repair. If the anchor is correct outside an unknown set of at most k_{rep} cells, selective inheritance pays k_{rep}/n for estimating the repair cells, together with a $\sqrt{K}\,\ell/n$ penalty for identifying and auditing the repair region, rather than the K/n cost

of refitting every cell. Hence its risk is asymptotically smaller than that of full cellwise empirical refitting when $k_{\mathrm{rep}} + \sqrt{K}\,\ell = o(K)$. In this sense, selective inheritance converts a global re-estimation problem into a smaller residual repair problem whenever most of the anchor is correct.

5.2 Finite-sample guarantee for holdout-based selection

We now return to the procedure in Alg. 1 and analyze its holdout-based predictor selection.

Let M_{anc}, M_{pres}, and M_{rep} denote the numbers of candidate–test pairs for the anchor, preservation, and repair audits, respectively, as defined in App. D.1.

Theorem 5.2 (Observable finite-path selection guarantee)**.** *Conditional on the preprocessing and construction data, suppose that all candidate predictors and audit functions used in holdout evaluation are fixed before the independent holdouts are observed. Let $n_{\mathrm{ho}} := |S_{\mathrm{ho}}|$ and $m_{\mathrm{ho}} := |U_{\mathrm{ho}}|$. For any $\delta \in (0,1)$, define $\xi_{\mathrm{pres}} := \sqrt{2\log(8M_{\mathrm{pres}}/\delta)/m_{\mathrm{ho}}}$, $\xi_{\mathrm{anc}} := W\sqrt{2\log(8M_{\mathrm{anc}}/\delta)/n_{\mathrm{ho}}}$, and $\xi_{\mathrm{rep}} := W\sqrt{2\log(8M_{\mathrm{rep}}/\delta)/n_{\mathrm{ho}}}$. Then, with probability at least $1-\delta$ over the independent holdouts, the predictor $\hat{q}$ returned by Alg. 1 satisfies*

$$C_{\mathrm{obs}}(\hat{q}) \le \min_{q \in \mathcal{E}_{\hat{q}}} C_{\mathrm{obs}}(q) + 2\max\left\{\xi_{\mathrm{anc}} + \xi_{\mathrm{pres}}, \xi_{\mathrm{rep}}\right\}.$$

Interpretation & practical implication. The theorem shows that the selected predictor is within an explicit finite-sample allowance of the best eligible predictor on the selected path. This remains valid even though the holdouts are used for selection, because all candidate predictors and audit tests are fixed beforehand.

In practice, candidates and tests should therefore be fixed before holdout evaluation, and final selection should use uncertainty-adjusted rather than raw empirical violations. The allowance decreases at the inverse-square-root rate with holdout size and grows only logarithmically with the number of candidate–test pairs. The guarantee is for C_{obs}, not target calibration, which additionally requires the transport conditions below.

5.3 From observable audits to target validity

The observable selection guarantee does not determine whether target labels agree with their weighted source counterparts. The following proposition formalizes the obstruction.

Proposition 5.3 (No target guarantee without a transfer restriction)**.** *Let $A \subseteq \mathcal{X}$ have target mass $\gamma > 0$, and suppose the target audit class contains $\mathbf{1}_A$. There exist two target data-generating processes with the same source labeled law, target covariate law, and paired model outputs such that, for every predictor produced from the available data, the following inequality holds under at least one of these processes:*

$$\left|\mathbb{E}_{P_{\mathrm{tgt}}}[\mathbf{1}_A(X)(Y - q(X))]\right| \ge \gamma/2.$$

Interpretation & practical implication. This is an information limit. Two target data-generating processes can have identical observable distributions of source labels, target covariates, and paired model scores while requiring incompatible calibrated predictions. No unlabeled-target method can distinguish them. Therefore, any target guarantee must state an explicit source-to-target restriction; without one, the observable guarantee in Thm. 5.2 is the valid claim supported by the available data.

Next, let the target calibration residual be $C_{\mathrm{tgt}}(q) := \max\{\sup_{a\in\mathcal{A}_{\mathrm{pres}}} |\mathbb{E}_{P_{\mathrm{tgt}}}[a(X)\{Y - q(X)\}]|, \sup_{b\in\mathcal{A}_{\mathrm{rep}}(q)} |\mathbb{E}_{P_{\mathrm{tgt}}}[b(X)\{Y - q(X)\}]|\}$. Using $Y - q = (Y - q^{\mathsf{a}}) + (q^{\mathsf{a}} - q)$, the preservation residual separates into an anchor residual and the observable deviation $V_{\mathrm{pres}}(q)$. Let β_{anc} and $\beta_{\mathrm{rep}}(q)$ bound, respectively, the source-to-target gaps in the anchor and repair residuals (see App. D.3 for the precise definitions). Then, the following inequality holds for any preservation and repair audit classes.

Proposition 5.4 (Target calibration decomposition via transfer gaps)**.** *Every predictor q satisfies*

$$C_{\mathrm{tgt}}(q) \le \max\{V_{\mathrm{anc}} + V_{\mathrm{pres}}(q) + \beta_{\mathrm{anc}},\ V_{\mathrm{rep}}(q) + \beta_{\mathrm{rep}}(q)\}.$$

Interpretation. The inequality is algebraically simple but identifies exactly what separates the observable guarantee from target calibration. The audited terms V_{anc}, $V_{\mathrm{pres}}(q)$, and $V_{\mathrm{rep}}(q)$ form the observable inheritance-or-repair criterion. The remaining terms are two source-to-target residual gaps. Thus, for the fixed audited classes, small observable violations imply small target violations

precisely to the extent that these two gaps are small. The decomposition makes the unverified part of a target claim explicit rather than hiding it inside a generic distribution-shift assumption.

In particular, if $P_{\mathrm{tgt}}(Y \mid X) = P_{\mathrm{src}}(Y \mid X)$ and $\hat{w} = dP^X_{\mathrm{tgt}}/dP^X_{\mathrm{src}}$ on the audited regions, change of measure gives $\beta_{\mathrm{anc}} = \beta_{\mathrm{rep}} = 0$ and $C_{\mathrm{tgt}}(q) \le C_{\mathrm{obs}}(q)$. If exact transport fails, the two-gap bound still holds, but the gaps cannot be identified from the available unlabeled target data.

Combining the guarantees. Together, the results separate two sources of uncertainty. Thm. 5.2 controls the finite-sample error from candidate selection, while Prop. 5.4 controls the additional error due to source-to-target transfer. We obtain

$$C_{\mathrm{tgt}}(\hat{q}) \le \min_{q \in \mathcal{E}_{\hat{\theta}}} C_{\mathrm{obs}}(q) + 2\max\{\xi_{\mathrm{anc}} + \xi_{\mathrm{pres}}, \xi_{\mathrm{rep}}\} + \max\{\beta_{\mathrm{anc}}, \beta_{\mathrm{rep}}(\hat{q})\}.$$

Prop. 5.3 shows why the final term cannot be eliminated from unlabeled target data alone. Thus, the unconditional finite-sample guarantee is for C_{obs}; target calibration additionally depends on the anchor and repair transfer gaps.

6 Empirical Illustration

We investigate whether A-PAIR can selectively inherit predecessor information and calibrate fine-tuned confidence without target labels on two real-world datasets, using `Qwen2.5-1.5B` as the predecessor and its instruction-tuned counterpart `Qwen2.5-1.5B-Instruct` as the fine-tuned model (Qwen et al., 2024). The full experimental details are given in App. G.

Data protocol. We use two multiple-choice benchmarks: MMLU-Pro (Wang et al., 2024), a knowledge and reasoning benchmark covering 14 broad subjects with 12,032 questions, and SciQ (Welbl et al., 2017), a four-choice science benchmark with 13,635 valid questions after preprocessing. For each question, let $\hat{j}$ be the answer selected by the fine-tuned model. We study calibration for the binary event that $\hat{j}$ is correct; both models' confidences refer to this same option. We split labeled source data into construction and independent audit sets, with cross-validation for A-PAIR's repair stopping coefficient performed within the construction set. Target labels are used only for final evaluation, after the anchor, stable/drift partition, candidate paths, and final predictor have been fixed.

Methods. For A-PAIR, we use Alg. 1, splitting the overall significance level $\delta = 0.05$ equally among the anchor, preservation, repair, and source-Brier audits ($\delta_{\mathrm{stage}} = 0.0125$). For the anchor and preservation audits, we compute simultaneous audit radii using studentized Bonferroni approximations. We compare against the raw fine-tuned score (Raw-FT), CPCS (Park et al., 2020), and DACA (Luo et al., 2025). CPCS transfers source calibration to the target distribution under covariate shift using importance weighting (App. G.2). DACA uses paired predictions from both models to calibrate confidence without correctness labels, making it a natural comparison for methods that retain access to the predecessor. All methods share the same final target evaluation sets.

We evaluate $\mathrm{TCE}(q) := \max_{h \in \mathcal{H}_{\mathrm{eval}}} \left|\widehat{\mathbb{E}}_{\mathrm{eval}}[h(X)(Y - q(X))]\right|$, an external empirical calibration metric over a common label-independent class $\mathcal{H}_{\mathrm{eval}}$ of group and score-bin tests, with the same moment-residual form as C_{tgt}. We also report Brier score; 10-bin expected calibration error (ECE); smooth calibration error (smCE), defined as $\sup_f |\mathbb{E}[f(q)(Y - q)]|$ over $[-1, 1]$-valued 1-Lipschitz functions (Błasiok et al., 2023); and area under the receiver operating characteristic curve (AUROC).

Empirical evidence of retained information. Before comparing calibration methods, we investigate whether the predecessor's confidence remains informative about correctness among examples with nearly the same fine-tuned confidence. Within 0.01-wide intervals of fine-tuned probability, the accuracy difference between examples with high versus low predecessor scores is 8.86 percentage points on MMLU-Pro and 0.79 on SciQ after reweighting the two groups to have the same distribution over fine-tuned-confidence intervals (App. G.3). Thus, predecessor confidence provides additional discrimination of correctness within narrow ranges of fine-tuned confidence, consistent with the predecessor retaining label-relevant resolution beyond the fine-tuned score.

Results on MMLU-Pro. For MMLU-Pro, we induce controlled subject-mixture shifts. In each scenario, one subject has mass 0.04 in the source and 0.16 in the target, with the remaining mass distributed uniformly across the other 13 subjects. For each subject, we run the experiment three times on different fixed subsets of the data, yielding 42 runs in total. We first average the three runs within each subject, then average equally across subjects.

Table 1: Mean target performance under label-restricted selection. Lower is better for TCE, Brier, ECE, and smCE; higher is better for AUROC. Bold indicates the best mean within each dataset.

	MMLU-Pro					SciQ				
Method	TCE	Brier	ECE	smCE	AUROC	TCE	Brier	ECE	smCE	AUROC
Raw-FT	0.2334	0.241779	0.2341	0.2335	0.7190	0.0083	0.013876	0.0094	0.0082	**0.9722**
CPCS	0.0271	0.181229	0.0378	0.0215	0.7190	**0.0042**	0.013298	**0.0074**	**0.0028**	0.9718
DACA	0.0237	0.180268	**0.0371**	0.0189	0.7193	0.0103	0.013331	0.0144	0.0101	0.9716
A-PAIR	**0.0207**	**0.178757**	0.0379	**0.0159**	**0.7228**	0.0061	**0.013276**	0.0084	0.0056	0.9603

Under these subject-mixture shifts, A-PAIR performs strongly overall relative to the baselines, with gains across both calibration and ranking measures and only a small exception on ECE. Importantly, this performance does not come from relying entirely on repair. Inheritance is used in all 42 runs, covers at least 10% of target examples in 40 runs, and has mean coverage of 18.6%. Thus, A-PAIR achieves competitive performance while using the intended preserve-or-repair structure: reusing predecessor information on selected regions and repairing the remainder.

Results on SciQ. For SciQ, we rank examples by the absolute logit difference between the fine-tuned and predecessor models $|\operatorname{logit}(q_{\mathrm{ft}}) - \operatorname{logit}(q_{\mathrm{pre}})|$. To construct the source–target shift, we form samples with different mixture proportions over the five discrepancy quintiles: 0.08 versus 0.32 for the highest-discrepancy quintile, and 0.23 versus 0.17 for each remaining quintile (source versus target; App. G.1). We report results over 30 fixed splits.

The method comparison in Tab. 1 and the diagnostic in App. G.3 distinguish retained information from realized calibration gains. Although the fine-tuned model is approximately 98% accurate on SciQ, predecessor confidence still separates the remaining errors: within narrow fine-tuned confidence ranges, weighted error rates are 1.145% and 0.352% for the low- and high-predecessor groups. Yet scalar temperature correction is already highly competitive (App. G.3): CPCS achieves lower mean TCE and smCE, while A-PAIR slightly improves Brier. Thus, the residual predecessor signal does not translate into consistent gains across metrics in this low-error regime, illustrating a method-selection boundary between settings where retained predecessor information can support selective inheritance and those where a simple score-only correction is already sufficient.

7 Discussion

This work shows that the model before fine-tuning can retain information lost after fine-tuning, and that reusing it selectively can make calibration more efficient.

The predecessor as calibration memory. Fine-tuning can miscalibrate confidence in two different ways. It can distort the scale of confidence, which score-only recalibration can address. It can also collapse distinctions that matter for correctness, which no transformation of the final score can undo. In the second case, the predecessor can act as a calibration memory that keeps the lost distinctions. Practically, this view also helps in choosing a method. Among examples with nearly the same fine-tuned confidence, one can check whether the predecessor's confidence still separates correct from incorrect answers. This check indicates in advance which type of miscalibration is dominant.

The role of selective inheritance. In practice, A-PAIR selects, among the candidates that pass the anchor audit, the one with the largest stable-region coverage. Coverage, however, is only a proxy. The real benefit of selective inheritance is that it removes regions where the anchor is already adequate from re-estimation, so that repair becomes a smaller residual problem. The predecessor also plays more than one role. Beyond serving as the reference to preserve, it helps define the subgroups that need repair. The predecessor is thus both an anchor to keep and a source of features that guide repair.

Scope and outlook. Our guarantees apply to an observable audit criterion; target calibration also depends on transfer gaps that unlabeled target data cannot identify. This limitation applies to any calibration method without target labels, and we make it explicit. Since earlier and updated models often coexist, as in preference tuning, quantization, and continual updates, inheriting useful calibration information across model updates is a natural next step.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, design or provide feedback on experiments, implement methods, and assist with translation. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results, and no other required-disclosure tasks are applicable to this work. Additionally, we used generative AI tools to suggest experimental parameters, search for and identify relevant literature, assist with LaTeX formatting, and edit the paper to improve readability through proofreading. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 72071–72095. Curran Associates, Inc., 2023. doi: 10.52202/075280-3154. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e4165c96702bac5f4962b70f3cf2f136-Abstract-Conference.html.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. doi: 10.48550/arXiv.2108.07258. URL https://arxiv.org/abs/2108.07258.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://academic.oup.com/book/26549.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, January 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2. URL https://journals.ametsoc.org/abstract/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Gabriele Campanella, Neeraj Kumar, Swaraj Nanda, Siddharth Singi, Eugene Fluder, Ricky Kwan, Silke Muehlstedt, Nicole Pfarr, Peter J. Schüffler, Ida Häggström, Noora Neittaanmäki, Levent M. Akyürek, Alina Basnet, Tamara Jamaspishvili, Michel R. Nasr, Matthew M. Croken, Fred R. Hirsch, Arielle Elkrief, Helena Yu, Orly Ardon, Gregory M. Goldgof, Meera Hameed, Jane Houldsworth, Maria Arcila, Thomas J. Fuchs, and Chad Vanderbilt. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. *Nature Medicine*, 31(9):3002–3010, 2025. doi: 10.1038/s41591-025-03780-x.

Fabrizio Carcillo, Andrea Dal Pozzolo, Yann-Aël Le Borgne, Olivier Caelen, Yannis Mazzer, and Gianluca Bontempi. SCARFF: A scalable framework for streaming credit card fraud detection with Spark. *Information Fusion*, 41:182–194, 2018. doi: 10.1016/j.inffus.2017.09.005.

Zhun Deng, Cynthia Dwork, and Linjun Zhang. HappyMap: A generalized multicalibration method. In Yael Tauman Kalai (ed.), *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 41:1–41:23, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-263-1. doi: 10.4230/LIPIcs.ITCS.2023.41. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2023.41.

Gianluca Detommaso, Martin Andres Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in LLMs. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10624–10641. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/detommaso24a.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11459–11492. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/globus-harris23a.html.

Parikshit Gopalan, Michael P Kim, Mihir A Singhal, and Shengjia Zhao. Low-degree multicalibration. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3193–3234. PMLR, 02–05 Jul 2022a. URL https://proceedings.mlr.press/v178/gopalan22a.html.

Parikshit Gopalan, Omer Reingold, Vatsal Sharan, and Udi Wieder. Multicalibrated partitions for importance weights. In Sanjoy Dasgupta and Nika Haghtalab (eds.), *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 408–435. PMLR, 29 Mar–01 Apr 2022b. URL https://proceedings.mlr.press/v167/gopalan22a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Nika Haghtalab, Michael Jordan, and Eric Zhao. A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 72464–72506. Curran Associates, Inc., 2023. doi: 10.52202/075280-3169. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e55edcdb01ac45c839a602f96e09fbcb-Paper-Conference.pdf.

Dutch Hansen, Siddartha Devic, Preetum Nakkiran, and Vatsal Sharan. When is multicalibration post-processing necessary? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 38383–38455. Curran Associates, Inc., 2024. doi: 10.52202/

079017-1213. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/43c6d9b0d11696f89c75157919a0334a-Paper-Conference.pdf.

Guande He, Jianfei Chen, and Jun Zhu. Preserving pre-trained features helps calibrate fine-tuned language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NI7StoWHJPT.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Emilien Joly and Gábor Lugosi. Robust estimation of U-statistics. *Stochastic Processes and their Applications*, 126(12):3760–3773, 2016. doi: 10.1016/j.spa.2016.04.021. URL https://doi.org/10.1016/j.spa.2016.04.021.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254. ACM, 2019. doi: 10.1145/3306618.3314287. URL https://dl.acm.org/doi/10.1145/3306618.3314287.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. Calibrated language model fine-tuning for in- and out-of-distribution data. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1326–1340, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.102. URL https://aclanthology.org/2020.emnlp-main.102/.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Beier Luo, Shuoyuan Wang, Sharon Li, and Hongxin Wei. Your pre-trained LLM is secretly an unsupervised confidence calibrator. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 163514–163543. Curran Associates, Inc., 2025. doi: 10.52202/085713-5458. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ef6b667990416111e3ebbeeba4a8f283-Abstract-Conference.html.

Zheda Mai, Arpita Chowdhury, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Vardaan Pahuja, Tanya Berger-Wolf, Song Gao, Charles Stewart, Yu Su, and Wei-Lun Chao. Fine-tuning is fine, if calibrated. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 136084–136119. Curran Associates, Inc., 2024. doi: 10.52202/079017-4323. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f573c36434796efe066d2f4cf3349e7f-Abstract-Conference.html.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.learningtheory.org/colt2009/papers/012.pdf.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, June 1973. URL https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Preetum Nakkiran, Arwen Bradley, Adam Goliński, Eugene Ndiaye, Michael Kirchhof, and Sinead Williamson. Trained on tokens, calibrated on concepts: The emergence of semantic calibration in LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/39b228b1a83232b14a1c81b0ed5a7a05-Abstract-Conference.html.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3219–3229. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/park20b.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1980. doi: 10.1002/9780470316481. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316481.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z. URL https://doi.org/10.1007/s10208-011-9099-z.

Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Transferable calibration with lower bias and variance in domain adaptation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19212–19223. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/df12ecd077efc8c23881028604dbb8cc-Paper.pdf.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

Jiayun Wu, Jiashuo Liu, Peng Cui, and Zhiwei Steven Wu. Bridging multicalibration and out-of-distribution generalization beyond covariate shift. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 73036–73078. Curran Associates, Inc., 2024. doi: 10.52202/079017-2325. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/859b6564b04959833fdf52ae6f726f84-Abstract-Conference.html.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. On the calibration of large language models and alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9778–9795, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.654. URL https://aclanthology.org/2023.findings-emnlp.654/.

Appendix Contents

A Related Work **15**

B Notation and Audit Setup **16**

B.1 Sample Splits, Candidate Families, and Frozen Objects 16

B.2 Audit Function Classes and Associated Quantities 16

C Information Retained by the Score Pair **18**

C.1 Resolution Lost Under Score Collapse (Proof of Prop. 2.1) 18

C.2 Anchor-Gap Separation on a Collapsed Cell (Lem. C.1 and Proof) 18

D Observable Guarantees and Target Validity **18**

D.1 Finite-Path Holdout Selection (Proof of Thm. 5.2) 18

D.2 Target Non-Identifiability (Proof of Prop. 5.3) 20

D.3 Transfer-Gap Decomposition and Exact Transport (Proof of Prop. 5.4) 20

D.4 Alternative Fresh-Audit Analysis (Prop. D.2 and Proof) 20

E Selective-Inheritance Theory **21**

E.1 Auxiliary Quadratic Audited Selector 22

E.2 Certificate for Joint Partition and Predictor Selection (Lem. E.2 and Proof) 24

E.3 Adaptation to an Unknown Repair Region (Lem. E.9 and Proof) 26

E.4 General Rate for a Finite Class of Audit Functions (Lem. E.10 and Proof) 28

E.5 Statistical Separation from Full Refitting (Thm. E.11 and Proof) 29

F Implementation Details **30**

F.1 Anchor Construction and Candidate Stable Regions 30

F.2 Density-Ratio Estimation 31

F.3 Frozen Objects, Source Screening, and Finite Classes of Audit Functions 31

G Experimental Details **32**

G.1 Models, Distribution Shifts, and Data Splits 32

G.2 Calibration Methods and Audit Protocol 32

G.3 Evaluation Metrics and Additional Diagnostics 33

A Related Work

We review two main lines of related work and highlight how our contributions differ from prior work in each.

Multicalibration. Multicalibration strengthens ordinary calibration by requiring calibration simultaneously across many, potentially overlapping subpopulations (Hébert-Johnson et al., 2018). Multiaccuracy relaxes this requirement to subgroup residual constraints and supports black-box post-processing of an existing predictor (Kim et al., 2019), while low-degree and smooth multicali-

bration replace exact conditioning on the prediction score with restricted classes of score-dependent tests, improving statistical and computational tractability (Gopalan et al., 2022a). Algorithmic work further connects these moment conditions to boosting, squared-error regression, and more general multi-objective or generalized-multicalibration frameworks (Globus-Harris et al., 2023; Haghtalab et al., 2023; Deng et al., 2023). Recent work has applied multicalibration directly to confidence scoring in language models and studied when multicalibration post-processing yields empirical benefits beyond ordinary calibration (Detommaso et al., 2024; Hansen et al., 2024). A-PAIR departs from this global post-processing view. Rather than recalibrating the entire predictor with a single multicalibration procedure, it first treats predecessor-derived confidence as information that may be preserved, audits whether such inheritance is supported on candidate stable regions, and confines label-based multicalibration updates to the remaining drift region. Our contribution is therefore an audited inheritance-or-repair architecture in which multicalibration serves as localized repair rather than global refitting. This structure is what allows A-PAIR to reuse information retained across fine-tuning while concentrating label-based estimation on the regions that require correction.

Calibration after fine-tuning. Fine-tuning and post-training can alter calibration in structured ways even when predictive capability is largely retained. OpenAI (2023) reports that the pretrained GPT-4 model is well calibrated, but that post-training substantially degrades its calibration. Zhu et al. (2023) and Mai et al. (2024) document related effects of alignment and fine-tuning on calibration. Nakkiran et al. (2026) study semantic calibration in language models, linking calibration in base models to local loss optimality and showing that instruction tuning and chain-of-thought reasoning can disrupt it. Their setting concerns distributions over semantic answer classes induced by sampling, whereas we calibrate the correctness of a fixed answer selected by the fine-tuned model. Two closely related methods are CPCS and DACA. CPCS transfers source calibration to an unlabeled target distribution through importance weighting under covariate shift (Park et al., 2020), whereas DACA uses the pretrained model as an unsupervised confidence reference for the post-trained model by aligning confidence on examples where the two models agree (Luo et al., 2025). A-PAIR combines these ingredients in a fundamentally different way: source-to-target importance weighting supports auditing, while predecessor information is inherited only in regions where the anchor audit supports reuse; the remaining regions are repaired using labeled source data. A-PAIR thus treats predecessor reuse as an audited, region-selective decision rather than a global calibration target.

B Notation and Audit Setup

B.1 Sample Splits, Candidate Families, and Frozen Objects

We work with binary labels $Y \in \{0,1\}$, labeled source distribution P_{src}, and target distribution P_{tgt}, from which the procedure observes only covariates. For each $\theta_{\mathrm{anc}} \in \Theta$, define the anchor $q_{\mathrm{anc},\theta_{\mathrm{anc}}}$ and indicators $s_{\theta_{\mathrm{anc}}}$ and $d_{\theta_{\mathrm{anc}}} := 1 - s_{\theta_{\mathrm{anc}}}$. The all-repair candidate θ_{rep} has $s_{\theta_{\mathrm{rep}}} \equiv 0$ and $d_{\theta_{\mathrm{rep}}} \equiv 1$. Every candidate has a bounded initializer $q^{(0)}_{\theta_{\mathrm{anc}}} \in [0,1]$. The initializer and path thresholds are fixed and are not separately compared on the holdouts. Conditional on preprocessing and construction data, every path $\mathcal{Q}_{\theta_{\mathrm{anc}}} = \{q^{(0)}_{\theta_{\mathrm{anc}}}, \ldots, q^{(T_{\theta_{\mathrm{anc}}})}_{\theta_{\mathrm{anc}}}\}$, weight, bin boundary, and audit function is fixed before the source and target holdouts are inspected. Selection is handled by one simultaneous event over all pairs $(\theta_{\mathrm{anc}}, q)$ with $\theta_{\mathrm{anc}} \in \Theta \cup \{\theta_{\mathrm{rep}}\}$ and $q \in \mathcal{Q}_{\theta_{\mathrm{anc}}}$.

B.2 Audit Function Classes and Associated Quantities

The quantities V_{anc}, V_{pres}, and V_{rep} in the main text correspond to the finite-class population quantities defined below, with candidate indices suppressed.

Let $\mathcal{G}_{\mathrm{pres}}$ and $\mathcal{G}_{\mathrm{rep}}$ contain functions bounded by one, let $\Phi_{\le k}$ be the bounded low-degree anchor-score class, and let $\mathcal{B}$ be the fixed current-score bin family. For each candidate $\theta_{\mathrm{anc}} \in \Theta$, $\mathcal{A}^{\mathrm{pop}}_{\mathrm{pres}}(\theta_{\mathrm{anc}})$ consists of tests $s_{\theta_{\mathrm{anc}}} g \phi(q_{\mathrm{anc},\theta_{\mathrm{anc}}})$ with $g \in \mathcal{G}_{\mathrm{pres}}$ and $\phi \in \Phi_{\le k}$. These tests restrict a subgroup to the stable region and localize it along the anchor score. The repair class $\mathcal{A}^{\mathrm{pop}}_{\mathrm{rep}}(q, \theta_{\mathrm{anc}})$ consists of tests $d_{\theta_{\mathrm{anc}}} g \mathbf{1}\{q \in I\}$ with $g \in \mathcal{G}_{\mathrm{rep}}$ and $I \in \mathcal{B}$, restricting a subgroup to the repair region and a current-score bin. For θ_{rep}, the preservation class is absent and the repair definition applies with

Table 2: Data splits and their roles. The "const" samples are used for construction, while the "ho" samples are independent audit holdouts. Target labels are accessed only for final evaluation.

Split	Domain and label access	Main uses
S_{const}	source, labeled	Source side of density-ratio estimation; internal cross-validation; repair-path construction.
U_{const}	target, unlabeled	Target side of density-ratio estimation; anchor-map fitting; preservation-path construction.
S_{ho}	source, labeled	Anchor and repair audits; paired source-Brier screen.
U_{ho}	target, unlabeled	Stable-region coverage and preservation-deviation audit.
Target evaluation set	target, labels revealed only at evaluation	Performance evaluation of the already selected predictor.

$d_{\theta_{\text{rep}}} \equiv 1$. The finite classes used in practice satisfy $\mathcal{A}^{\text{fin}}_{\text{pres}}(\theta_{\text{anc}}) \subseteq \mathcal{A}^{\text{pop}}_{\text{pres}}(\theta_{\text{anc}})$ and $\mathcal{A}^{\text{fin}}_{\text{rep}}(q, \theta_{\text{anc}}) \subseteq \mathcal{A}^{\text{pop}}_{\text{rep}}(q, \theta_{\text{anc}})$.

For a bounded test class $\mathcal{A}$, write $\text{Cal}_P(q; \mathcal{A}) := \sup_{a \in \mathcal{A}} |\mathbb{E}_P[a(X)(Y - q(X))]|$. For each candidate $\theta_{\text{anc}} \in \Theta$, the generic population functionals are

$$
\begin{aligned}
V^{\text{pop}}_{\text{anc}}(\theta_{\text{anc}}) &:= \sup_{a \in \mathcal{A}^{\text{pop}}_{\text{pres}}(\theta_{\text{anc}})} \left|\mathbb{E}_{P_{\text{src}}}[\hat{w}(X)a(X)(Y - q_{\text{anc},\theta_{\text{anc}}}(X))]\right|, \\
V^{\text{pop}}_{\text{pres}}(q; \theta_{\text{anc}}) &:= \sup_{a \in \mathcal{A}^{\text{pop}}_{\text{pres}}(\theta_{\text{anc}})} \left|\mathbb{E}_{P^X_{\text{tgt}}}[a(X)(q(X) - q_{\text{anc},\theta_{\text{anc}}}(X))]\right|, \\
V^{\text{pop}}_{\text{rep}}(q; \theta_{\text{anc}}) &:= \sup_{b \in \mathcal{A}^{\text{pop}}_{\text{rep}}(q, \theta_{\text{anc}})} \left|\mathbb{E}_{P_{\text{src}}}[\hat{w}(X)b(X)(Y - q(X))]\right|.
\end{aligned}
$$

The quantities for finite audit classes $V^{\text{fin}}_{\text{anc}}(\theta_{\text{anc}})$, $V^{\text{fin}}_{\text{pres}}(q; \theta_{\text{anc}})$, and $V^{\text{fin}}_{\text{rep}}(q; \theta_{\text{anc}})$ use the same moments, with each supremum restricted to its finite class of audit functions. Thus, the change from pop to fin changes only the test class, not the distribution or residual.

We recompute the predictor-dependent audits on independent holdouts. The preservation audit $\widehat{V}_{\text{pres,ho}}(q; \theta_{\text{anc}})$ replaces $\mathbb{E}_{P^X_{\text{tgt}}}$ in $V^{\text{fin}}_{\text{pres}}$ by the sample average $\widehat{\mathbb{E}}_{U_{\text{ho}}}$. The repair audit $\widehat{V}_{\text{rep,ho}}(q; \theta_{\text{anc}})$ replaces $\mathbb{E}_{P_{\text{src}}}$ in $V^{\text{fin}}_{\text{rep}}$ by $\widehat{\mathbb{E}}_{S_{\text{ho}}}$, retaining the weight $\hat{w}$ and the same finite class of repair functions.

For each candidate $\theta_{\text{anc}} \in \Theta$, define $C^{\text{pop}}_{\text{obs}}(q; \theta_{\text{anc}}) := \max\{V^{\text{pop}}_{\text{anc}}(\theta_{\text{anc}}) + V^{\text{pop}}_{\text{pres}}(q; \theta_{\text{anc}}), V^{\text{pop}}_{\text{rep}}(q; \theta_{\text{anc}})\}$ and define $C^{\text{fin}}_{\text{obs}}$ analogously using the finite audit classes. The empirical upper guarantee $\widehat{C}^{\text{fin}}_{\text{obs}}(q; \theta_{\text{anc}})$ adds the simultaneous radii to the corresponding holdout quantities. For the all-repair candidate, we explicitly set $V_{\text{anc}} = V_{\text{pres}}(q) = 0$ and use the repair-only definitions $C_{\text{obs}}(q) = V_{\text{rep}}(q)$ and $\widehat{C}_{\text{obs}}(q) = \widehat{V}_{\text{rep,ho}}(q) + \xi_{\text{rep}}$. The same repair-only convention is used for both the population and empirical guarantees.

For the generic population classes, let $C^{\text{pop}}_{\text{tgt,pres}}(q; \theta_{\text{anc}}) := \text{Cal}_{P_{\text{tgt}}}(q; \mathcal{A}^{\text{pop}}_{\text{pres}}(\theta_{\text{anc}}))$ and $C^{\text{pop}}_{\text{tgt,rep}}(q; \theta_{\text{anc}}) := \text{Cal}_{P_{\text{tgt}}}(q; \mathcal{A}^{\text{pop}}_{\text{rep}}(q, \theta_{\text{anc}}))$. Their maximum is $C^{\text{pop}}_{\text{tgt}}(q; \theta_{\text{anc}})$; replacing both classes by their finite subclasses gives $C^{\text{fin}}_{\text{tgt}}(q; \theta_{\text{anc}})$. For the all-repair candidate, C_{tgt} is the repair side alone. The finite-sample theorem concerns only the finite functionals, whereas the proof below establishes the deterministic two-gap inequality for the generic population classes. The statement for finite audit classes in the main text follows immediately by restricting the suprema to the classes used in the implementation.

C Information Retained by the Score Pair

C.1 Resolution Lost Under Score Collapse (Proof of Prop. 2.1)

Projection identity. The proof uses the following L_2 projection identity. If state variables S_1 and S_2 satisfy $\sigma(S_1) \subseteq \sigma(S_2)$, then

$$\inf_f \mathbb{E}[(Y - f(S_1))^2] - \inf_g \mathbb{E}[(Y - g(S_2))^2] = \mathbb{E}\left[\big(\mathbb{E}[Y \mid S_2] - \mathbb{E}[Y \mid S_1]\big)^2\right].$$

For any state variable S, conditional expectation is the L_2 projection of Y onto the S-measurable functions. Because $\mathbb{E}[\mathbb{E}[Y \mid S_2] \mid S_1] = \mathbb{E}[Y \mid S_1]$, the decomposition $Y - \mathbb{E}[Y \mid S_1] = Y - \mathbb{E}[Y \mid S_2] + \mathbb{E}[Y \mid S_2] - \mathbb{E}[Y \mid S_1]$ has orthogonal summands. Taking squared norms gives the identity.

Proof of Prop. 2.1. Fix a blind correction T and put $c := T(v)$. Define $R_T(u, w)$ to equal $\bar{\eta}_i$ when $w \in J_i$ for $i = 0, 1$, and to equal $T(u)$ when $w \notin J_0 \cup J_1$. The separation assumptions imply that R_T outputs $\bar{\eta}_i$ on C_i and agrees with $T(z_{\mathrm{ft}})$ elsewhere. Conditional centering at each cell mean gives $\mathbb{E}\left[\mathbf{1}_{C_i}\{(Y - c)^2 - (Y - \bar{\eta}_i)^2\}\right] = p_i(c - \bar{\eta}_i)^2$. The sum is minimized over c at $(p_0\bar{\eta}_0 + p_1\bar{\eta}_1)/(p_0 + p_1)$, with minimum $\frac{p_0 p_1}{p_0 + p_1}(\bar{\eta}_0 - \bar{\eta}_1)^2$. The infimum over pair-aware rules is no larger than the risk of R_T. This argument uses only cell means and does not require $\eta(x)$ to be constant within either cell. □

C.2 Anchor-Gap Separation on a Collapsed Cell (Lem. C.1 and Proof)

The structural separation requires the fitted anchor, rather than only the predecessor logit, to preserve a nonzero gap.

Lemma C.1 (Pair-aware guarantee separation)**.** *Let $C = C_0 \dot{\cup} C_1$ lie in the candidate stable region, with masses $p_i > 0$, cell means $\bar{\eta}_i$, and a collapsed fine-tuned logit on C. Suppose $q^{\mathsf{a}} = u_i$ on C_i and $\Delta_{\mathrm{anc}} := |u_1 - u_0| > 0$. If the audit class contains $h_0 = \mathbf{1}_C$ and $h_1 = \mathbf{1}_C q^{\mathsf{a}}$, then every score-only rule satisfies*

$$\max_{\ell=0,1} |\mathbb{E}[h_\ell(X)(Y - T(z_{\mathrm{ft}}(X)))]| \geq \frac{\Delta_{\mathrm{anc}}}{2} \frac{p_0 p_1}{p_0 + p_1} |\bar{\eta}_0 - \bar{\eta}_1|.$$

Every predictor q also satisfies

$$\max_{\ell=0,1} |\mathbb{E}[h_\ell(Y - q)]| \leq \max_{\ell=0,1} |\mathbb{E}[h_\ell(Y - q^{\mathsf{a}})]| + \max_{\ell=0,1} |\mathbb{E}[h_\ell(q - q^{\mathsf{a}})]|.$$

Proof. For a score-only rule, write $R_i := p_i(\bar{\eta}_i - T(v))$ and $M_\ell := \mathbb{E}[h_\ell(Y - T(v))]$. Then $M_0 = R_0 + R_1$ and $M_1 = u_0 R_0 + u_1 R_1$. Inverting this two-by-two system yields $\max_i |R_i| \leq 2 \max_\ell |M_\ell| / \Delta_{\mathrm{anc}}$. Minimizing $\max_i |R_i|$ over the shared value $T(v)$ gives $p_0 p_1 |\bar{\eta}_0 - \bar{\eta}_1| / (p_0 + p_1)$. The second inequality follows by adding and subtracting q^{a}. □

Finite-sample implication. For a finite candidate set of size M independent of an unlabeled target holdout of size m, the two empirical anchor-deviation moments are simultaneously accurate within $\sqrt{\frac{2\log(4M/\delta)}{m}}$ with probability at least $1 - \delta$. After inversion, an ε piecewise-residual remainder therefore requires $m \gtrsim \Delta_{\mathrm{anc}}^{-2} \varepsilon^{-2} \log(M/\delta)$.

The condition $\Delta_{\mathrm{anc}} > 0$ is essential. The identity and positive-slope affine maps used by the method preserve any nonzero predecessor gap under alignment.

D Observable Guarantees and Target Validity

D.1 Finite-Path Holdout Selection (Proof of Thm. 5.2)

Simultaneous audit radii. For counting purposes only, write $\bar{\Theta} := \Theta \cup \{\theta_{\mathrm{rep}}\}$, where Θ indexes the candidate anchors and stable regions. Let $N_{\mathrm{pres}} := \max_{\theta_{\mathrm{anc}} \in \Theta} |\mathcal{A}^{\mathrm{fin}}_{\mathrm{pres}}(\theta_{\mathrm{anc}})|$, $N_{\mathrm{rep}} := \max_{\theta_{\mathrm{anc}} \in \bar{\Theta},\, q \in \mathcal{Q}_{\theta_{\mathrm{anc}}}} |\mathcal{A}^{\mathrm{fin}}_{\mathrm{rep}}(q, \theta_{\mathrm{anc}})|$, and $M_Q := \sum_{\theta_{\mathrm{anc}} \in \bar{\Theta}} |\mathcal{Q}_{\theta_{\mathrm{anc}}}|$. Then $M_{\mathrm{anc}} := N_{\mathrm{pres}} |\Theta|$,

$M_{\mathrm{pres}} := N_{\mathrm{pres}} M_Q$, and $M_{\mathrm{rep}} := N_{\mathrm{rep}} M_Q$ bound the anchor, preservation, and repair candidate–test pairs compared on the same holdouts. Using the same M_Q in the preservation radius is conservative because it also counts predictors from the all-repair candidate, which has no preservation tests. The initializer and path thresholds are fixed and therefore do not contribute additional candidate factors. For target and source holdout sizes m_{ho} and n_{ho}, a Hoeffding instantiation (Boucheron et al., 2013, Sec. 2.6) sets the preservation radius to $\xi_{\mathrm{pres}} := \sqrt{2\log(8M_{\mathrm{pres}}/\delta)/m_{\mathrm{ho}}}$. The weighted source audits use $\xi_{\mathrm{anc}} := W\sqrt{2\log(8M_{\mathrm{anc}}/\delta)/n_{\mathrm{ho}}}$ for the anchor and $\xi_{\mathrm{rep}} := W\sqrt{2\log(8M_{\mathrm{rep}}/\delta)/n_{\mathrm{ho}}}$ for repair. The anchor radius ξ_{anc} is common across candidates. Valid simultaneous empirical-Bernstein radii may replace these Hoeffding radii; see, e.g., Maurer & Pontil (2009).

Paired source-Brier screen. For a tagged candidate $q \in \mathcal{Q}_{\theta_{\mathrm{anc}}}$, let $\Delta_{\mathrm{src}}(q;\theta_{\mathrm{anc}}) := \mathsf{Risk}_{P_{\mathrm{src}}}(q) - \mathsf{Risk}_{P_{\mathrm{src}}}(q^{(0)}_{\theta_{\mathrm{anc}}})$ be its population Brier difference from step zero. Its holdout counterpart, $\widehat{\Delta}_{\mathrm{src,ho}}(q;\theta_{\mathrm{anc}})$, is the average of $(Y-q(X))^2 - (Y - q^{(0)}_{\theta_{\mathrm{anc}}}(X))^2$ over S_{ho}. This is the quantity denoted by $\widehat{\Delta}_{\mathrm{src}}(q;\theta)$ in the main text, with $\theta = \theta_{\mathrm{anc}}$. Let M_{src} be the number of tagged candidate–step-zero comparisons included in the simultaneous family. A one-sided Hoeffding radius is $\xi_{\mathrm{src}} := \sqrt{\frac{2\log(4M_{\mathrm{src}}/\delta)}{n_{\mathrm{ho}}}}$. The upper confidence bound is $U_{\mathrm{src}}(q;\theta_{\mathrm{anc}}) := \widehat{\Delta}_{\mathrm{src,ho}}(q;\theta_{\mathrm{anc}}) + \xi_{\mathrm{src}}$. We always retain the initializer and retain other candidates only when this bound is nonpositive:

$$\mathcal{E}_{\theta_{\mathrm{anc}}} := \left\{q^{(0)}_{\theta_{\mathrm{anc}}}\right\} \cup \left\{q \in \mathcal{Q}_{\theta_{\mathrm{anc}}} \setminus \left\{q^{(0)}_{\theta_{\mathrm{anc}}}\right\} : U_{\mathrm{src}}(q;\theta_{\mathrm{anc}}) \le 0\right\}.$$

Each paired Brier difference lies in $[-1,1]$. Hence one-sided Hoeffding concentration and a union bound imply, with probability at least $1-\delta/4$, that $\Delta_{\mathrm{src}}(q;\theta_{\mathrm{anc}}) \le U_{\mathrm{src}}(q;\theta_{\mathrm{anc}})$ simultaneously over the tagged family. On this event, every candidate retained by the source-Brier screen satisfies $\mathsf{Risk}_{P_{\mathrm{src}}}(q) \le \mathsf{Risk}_{P_{\mathrm{src}}}(q^{(0)}_{\theta_{\mathrm{anc}}})$. Any valid simultaneous one-sided empirical-Bernstein radius may replace ξ_{src}.

Proof of Thm. 5.2. Condition on the preprocessing and construction data. The full tagged family of pairs $(\theta_{\mathrm{anc}}, q)$, with $\theta_{\mathrm{anc}} \in \Theta \cup \{\theta_{\mathrm{rep}}\}$ and $q \in \mathcal{Q}_{\theta_{\mathrm{anc}}}$, is then fixed before the holdouts are drawn, as are all anchors, indicators, weights, bin boundaries, and finite classes of audit functions.

For a candidate $(q, \theta_{\mathrm{anc}})$ with $\theta_{\mathrm{anc}} \in \Theta$ and preservation test a, the variable $a(X)(q(X) - q_{\mathrm{anc},\theta_{\mathrm{anc}}}(X))$ lies in $[-1,1]$. The weighted anchor and repair residuals lie in $[-W,W]$. Hoeffding's inequality and a union bound therefore give, simultaneously over the candidates in Θ,

$$\left|\widehat{V}_{\mathrm{pres,ho}}(q;\theta_{\mathrm{anc}}) - V^{\mathrm{fin}}_{\mathrm{pres}}(q;\theta_{\mathrm{anc}})\right| \le \xi_{\mathrm{pres}},$$
$$\left|\widehat{V}_{\mathrm{anc}}(\theta_{\mathrm{anc}}) - V^{\mathrm{fin}}_{\mathrm{anc}}(\theta_{\mathrm{anc}})\right| \le \xi_{\mathrm{anc}},$$
$$\left|\widehat{V}_{\mathrm{rep,ho}}(q;\theta_{\mathrm{anc}}) - V^{\mathrm{fin}}_{\mathrm{rep}}(q;\theta_{\mathrm{anc}})\right| \le \xi_{\mathrm{rep}}.$$

The repair inequality also holds for the all-repair candidate. The union of these events has probability at least $1-\delta$. Adding the first two inequalities and taking a maximum with the third proves the simultaneous upper guarantee, while the all-repair candidate uses only the repair inequality. The argument is unchanged when valid simultaneous empirical-Bernstein radii are substituted.

Uniformly over the family, $C^{\mathrm{fin}}_{\mathrm{obs}}(q;\theta_{\mathrm{anc}}) \le \widehat{C}^{\mathrm{fin}}_{\mathrm{obs}}(q;\theta_{\mathrm{anc}})$, while

$$\widehat{C}^{\mathrm{fin}}_{\mathrm{obs}}(q;\theta_{\mathrm{anc}}) \le C^{\mathrm{fin}}_{\mathrm{obs}}(q;\theta_{\mathrm{anc}}) + 2\max\{\xi_{\mathrm{anc}} + \xi_{\mathrm{pres}}, \xi_{\mathrm{rep}}\}.$$

Applying these inequalities to the empirical minimizer and a population minimizer gives the oracle inequality. □

Lemma D.1 (Finite empirical path length)**.** *Each accepted preservation update decreases $\widehat{L}$ by at least γ^2_{pres}, and each accepted repair update decreases it by at least $\gamma^2_{\mathrm{rep}}/W$. Hence*

$$T_\theta \le \left\lceil \frac{\widehat{L}(q^{(0)}_\theta)}{\min\{\gamma^2_{\mathrm{pres}}, \gamma^2_{\mathrm{rep}}/W\}} \right\rceil \le \left\lceil \frac{1+W}{\min\{\gamma^2_{\mathrm{pres}}, \gamma^2_{\mathrm{rep}}/W\}} \right\rceil.$$

With the iteration cap T in the main text, the number of accepted updates also satisfies $T_\theta \le T$.

D.2 Target Non-Identifiability (Proof of Prop. 5.3)

Proof of Prop. 5.3. Construct two target worlds with the same target covariate law, source joint law, and deterministic model outputs. Outside A, let their target label conditionals agree. On A, set $Y = 0$ almost surely in the first world and $Y = 1$ almost surely in the second. Target labels are unobserved, so the complete observable law is identical.

For any realized predictor q, the two violations are $V_0 = \mathbb{E}[\mathbf{1}_A(X)q(X)]$ and $V_1 = \mathbb{E}[\mathbf{1}_A(X)(1 - q(X))]$. Their sum is $P^X_{\mathrm{tgt}}(A) = \gamma$, hence $\max\{V_0, V_1\} \geq \gamma/2$. The argument is pointwise in the procedure's randomness. □

Specialization to anchor validity. To specialize the target non-identifiability construction to anchor validity, fix the covariate law, logits, map, and stable region S. In the first world, set $\mathbb{P}(Y = 1 \mid X = x) = q^{\mathsf{a}}(x)$ on S. Partition S as $E_- := S \cap \{q^{\mathsf{a}} \leq 1/2\}$ and $E_+ := S \cap \{q^{\mathsf{a}} > 1/2\}$. At least one set has positive target mass. If $P^X_{\mathrm{tgt}}(E_-) > 0$, choose a positive-mass measurable subset $E \subseteq E_-$ and set the second conditional mean to $\eta_2(x) = q^{\mathsf{a}}(x) + \delta$ on E, where $0 < \delta \leq 1/2$. Otherwise choose a positive-mass measurable subset $E \subseteq E_+$ and set $\eta_2(x) = q^{\mathsf{a}}(x) - \delta$ on E, again with $0 < \delta \leq 1/2$. Outside E, set $\eta_2 = q^{\mathsf{a}}$ on S, and keep both worlds identical outside S. The two worlds have identical covariate laws, logits, anchors, and indicators, but their anchor residuals differ on the preserve test $s_\tau \mathbf{1}_E$. Hence target anchor validity is not identified, including when q^{a} equals zero or one on the stable region.

D.3 Transfer-Gap Decomposition and Exact Transport (Proof of Prop. 5.4)

Definition of the two transfer gaps. For the preservation and repair classes in App. B, define

$$\begin{aligned}
\beta_{\mathrm{anc}}(\theta_{\mathrm{anc}}) := \sup_{a \in \mathcal{A}^{\mathrm{pop}}_{\mathrm{pres}}(\theta_{\mathrm{anc}})} \Big| & \mathbb{E}_{P_{\mathrm{tgt}}}[a(X)(Y - q_{\mathrm{anc},\theta_{\mathrm{anc}}}(X))] \\
& - \mathbb{E}_{P_{\mathrm{src}}}[\hat{w}(X)a(X)(Y - q_{\mathrm{anc},\theta_{\mathrm{anc}}}(X))]\Big|, \\
\beta_{\mathrm{rep}}(q; \theta_{\mathrm{anc}}) := \sup_{b \in \mathcal{A}^{\mathrm{pop}}_{\mathrm{rep}}(q,\theta_{\mathrm{anc}})} \Big| & \mathbb{E}_{P_{\mathrm{tgt}}}[b(X)(Y - q(X))] \\
& - \mathbb{E}_{P_{\mathrm{src}}}[\hat{w}(X)b(X)(Y - q(X))]\Big|.
\end{aligned}$$

Proof of Prop. 5.4. Fix θ_{anc}. For a preservation test a, adding and subtracting its anchor gives

$$\mathbb{E}_{P_{\mathrm{tgt}}}[a(X)(Y - q(X))] = \mathbb{E}_{P_{\mathrm{tgt}}}[a(X)(Y - q_{\mathrm{anc},\theta_{\mathrm{anc}}}(X))] + \mathbb{E}_{P^X_{\mathrm{tgt}}}[a(X)(q_{\mathrm{anc},\theta_{\mathrm{anc}}}(X) - q(X))].$$

Add and subtract the weighted source anchor residual, take absolute values, and then take the supremum to obtain

$$C^{\mathrm{pop}}_{\mathrm{tgt,pres}}(q; \theta_{\mathrm{anc}}) \leq V^{\mathrm{pop}}_{\mathrm{anc}}(\theta_{\mathrm{anc}}) + V^{\mathrm{pop}}_{\mathrm{pres}}(q; \theta_{\mathrm{anc}}) + \beta_{\mathrm{anc}}(\theta_{\mathrm{anc}}).$$

For a repair test b, adding and subtracting the weighted source repair residual gives

$$C^{\mathrm{pop}}_{\mathrm{tgt,rep}}(q; \theta_{\mathrm{anc}}) \leq V^{\mathrm{pop}}_{\mathrm{rep}}(q; \theta_{\mathrm{anc}}) + \beta_{\mathrm{rep}}(q; \theta_{\mathrm{anc}}).$$

Taking the maximum proves the proposition. □

Exact covariate transport. Fix θ_{anc}. Under the exact covariate transport condition on the stable and drift regions stated in Sec. 5.3, change of measure gives, for every audited residual $r(X, Y)$, $\mathbb{E}_{P_{\mathrm{tgt}}}[r(X, Y)] = \mathbb{E}_{P_{\mathrm{src}}}[\hat{w}(X)r(X, Y)]$. Thus $\beta_{\mathrm{anc}}(\theta_{\mathrm{anc}}) = \beta_{\mathrm{rep}}(q; \theta_{\mathrm{anc}}) = 0$. The generic population statement follows from Prop. 5.4. Applying the same identity to the finite classes of audit functions and then using the simultaneous holdout bounds gives the empirical guarantee.

D.4 Alternative Fresh-Audit Analysis (Prop. D.2 and Proof)

Population potential and descent. Fix θ_{anc}, and write s_τ and q^{a} for its indicator and anchor. The fresh-audit process uses the normalized updates in Alg. 1, but draws a new audit block at each round and returns q_{stop} rather than a holdout-selected predictor. Condition on the preprocessing

data, so s_τ, q^{a}, and $\hat{w}$ are fixed. The practical path uses one fixed construction sample, so its descent calculation is empirical and exact on that sample. The fresh-audit process is a separate theoretical analogue: it draws a new audit block at each round, pays for estimation error when converting the selected empirical correlation into population descent, and telescopes the resulting population potential $L(q) := L_{\mathrm{pres}}(q) + L_{\mathrm{rep}}(q)$. Here $L_{\mathrm{pres}}(q) := \mathbb{E}_{P^X_{\mathrm{tgt}}}[s_\tau(X)(q(X) - q^{\mathsf{a}}(X))^2]$ measures departure from the anchor on the stable region, and $L_{\mathrm{rep}}(q) := \mathbb{E}_{P_{\mathrm{src}}}[\hat{w}(X)d_\tau(X)(Y - q(X))^2]$ is the weighted source Brier loss on the repair region. For supported tests a and b, let $c_{\mathrm{pres}}(a; q) := \mathbb{E}_{P^X_{\mathrm{tgt}}}[a(X)(q(X) - q^{\mathsf{a}}(X))]$ and $c_{\mathrm{rep}}(b; q) := \mathbb{E}_{P_{\mathrm{src}}}[\hat{w}(X)b(X)(Y - q(X))]$. Hats denote the corresponding correlations on the current fresh audit block.

For the preservation update $q^+ = \mathrm{proj}_{[0,1]}(q - \widehat{c}_{\mathrm{pres}}a)$,

$$L_{\mathrm{pres}}(q) - L_{\mathrm{pres}}(q^+) \geq 2\widehat{c}_{\mathrm{pres}}c_{\mathrm{pres}} - \widehat{c}^2_{\mathrm{pres}}\mathbb{E}_{P^X_{\mathrm{tgt}}}[a(X)^2].$$

For the repair update $q^+ = \mathrm{proj}_{[0,1]}(q + \widehat{c}_{\mathrm{rep}}b/W)$,

$$L_{\mathrm{rep}}(q) - L_{\mathrm{rep}}(q^+) \geq \frac{2\widehat{c}_{\mathrm{rep}}c_{\mathrm{rep}}}{W} - \frac{\widehat{c}^2_{\mathrm{rep}}}{W^2}\mathbb{E}_{P_{\mathrm{src}}}[\hat{w}(X)d_\tau(X)b(X)^2].$$

If $|a|, |b| \leq 1$ and $|\widehat{c} - c| \leq |\widehat{c}|/4$ for the selected correlation, these decreases are at least $\widehat{c}^2_{\mathrm{pres}}/2$ and $\widehat{c}^2_{\mathrm{rep}}/(2W)$, respectively. Projection onto $[0, 1]$ cannot increase squared distance to the bounded anchor or to a binary label. Expanding the population squares before projection gives the two stated inequalities. The deviation event implies $\widehat{c}c \geq 3\widehat{c}^2/4$. Because $\mathbb{E}[a^2] \leq 1$ and $\mathbb{E}[\hat{w}d_\tau b^2] \leq W$, substitution gives the two lower bounds.

Proposition D.2 (Fresh-audit stopping)**.** *Suppose every fresh block estimates all signed preservation and repair correlations within $\gamma_{\mathrm{pres}}/4$ and $\gamma_{\mathrm{rep}}/4$, respectively. The fresh-audit path stops after at most*

$$T_{\max} = \left\lceil \frac{2(1+W)}{\min\{\gamma^2_{\mathrm{pres}}, \gamma^2_{\mathrm{rep}}/W\}} \right\rceil$$

rounds and returns

$$V^{\mathrm{fin}}_{\mathrm{pres}}(q_{\mathrm{stop}}; \theta_{\mathrm{anc}}) \leq \frac{5}{4}\gamma_{\mathrm{pres}}, \qquad V^{\mathrm{fin}}_{\mathrm{rep}}(q_{\mathrm{stop}}; \theta_{\mathrm{anc}}) \leq \frac{5}{4}\gamma_{\mathrm{rep}}.$$

Proof. Whenever a population violation over a finite audit class exceeds $5/4$ of its threshold, the corresponding empirical maximum exceeds the threshold. On any violating round, the absolute deviation bound also implies the relative condition in the preceding descent calculation. Hence

$$L(q^{(t)}) - L(q^{(t+1)}) \geq \frac{1}{2}\min\left\{\gamma^2_{\mathrm{pres}}, \frac{\gamma^2_{\mathrm{rep}}}{W}\right\}.$$

Because $L_{\mathrm{pres}}(q^{(0)}) \leq 1$, $L_{\mathrm{rep}}(q^{(0)}) \leq W$, and L is nonnegative, telescoping this single population potential gives $T_{\max}$. At stopping, the same deviation event yields the displayed $5/4$ bounds. □

Sample complexity. For fixed degree and audit-class complexities $\mathfrak{C}_{\mathrm{pres}}$ and $\mathfrak{C}_{\mathrm{rep}}$, per-round audit sizes are $\widetilde{O}\left(\frac{\mathfrak{C}_{\mathrm{pres}}}{\varepsilon^2}\right)$ and $\widetilde{O}\left(\frac{W^2\mathfrak{C}_{\mathrm{rep}}}{\varepsilon^2}\right)$. Together with $T_{\max} = O_W(\varepsilon^{-2})$, the total fresh-audit requirement is $\widetilde{O}(\varepsilon^{-4})$. This guarantees only the finite observable audits. A target-calibration statement additionally uses Prop. 5.4 or exact covariate transport.

E Selective-Inheritance Theory

This section develops the statistical theory of selective inheritance. The central phenomenon is adaptation: without being given the repair region, the procedure can exploit an anchor that is reliable on most of the state space and concentrate label-based estimation on the residual region that requires correction. As a result, the statistical complexity can scale with the complexity of the residual repair problem rather than that of globally re-estimating the full state space.

We formalize this phenomenon through a quadratic audited selector, establish joint selection guarantees over candidate partitions and predictors, derive adaptation rates for an unknown repair region, and obtain an explicit statistical separation from full refitting in a finite-cell model.

E.1 Auxiliary Quadratic Audited Selector

The auxiliary sample-split selector retains the anchor family, stable/drift partitions based on score discrepancies, and projected preserve-or-repair updates of A-PAIR, while selecting candidates θ and path predictors through a common quadratic certificate.

The common class of audit functions fixes the audit directions across candidate partitions, and anchor residual energy enters the selection score. The all-repair candidate is retained, and coverage may break exact ties. Unlike the coverage-first selection in the main text, this auxiliary selector jointly selects the candidate θ and predictor and permits a prespecified subset of path iterates to be saved. The adaptation analysis further uses anchor initialization on the stable region, $1/2$ on the repair region, and a sufficiently converged repair endpoint; its rate guarantee concerns this specialized construction.

For compact notation, write $a_\theta = q_{\mathrm{anc},\theta}$, with s_θ and $d_\theta = 1 - s_\theta$ as in App. B. No target labels enter the procedure.

Sample protocol and target geometry. Let $\mathcal{X}$ be any measurable input space and $Y \in \{0,1\}$. Assume

$$P_{\mathrm{tgt}}(Y \mid X) = P_{\mathrm{src}}(Y \mid X), \qquad w(x) = \frac{dP^X_{\mathrm{tgt}}}{dP^X_{\mathrm{src}}}(x) \in [0, W], \quad W \geq 1, \tag{E.1}$$

where the exact function w and an upper bound W are available. This is the standard covariate-shift change-of-measure regime: importance weighting transfers target expectations to weighted source expectations when the conditional label law is shared (Gopalan et al., 2022b; Wu et al., 2024). Under (E.1), the conditional success probability η defined in the main text is shared by the source and target distributions. We use the Brier risk defined in the main text. The general certificate lemma imposes no density lower bound or finite support assumption.

Independent preprocessing specifies a finite set of candidates θ, including the all-repair candidate with $s \equiv 0$. Independent source construction and audit samples have sizes n_0 and n_A, with total label budget $n = n_0 + n_A$. Construction data produce bounded predictors $q_i \in [0,1]$, where the tagged index $i = (\theta, t)$ ranges over a finite set $\mathcal{I}$. Every path includes its initializer $q^{(0)}_\theta$. In the general lemma we condition on all preprocessing and construction data. The saved family need not contain every internal numerical iteration; any thinning rule is fixed before audit data are opened.

Let $N, d \geq 1$ be deterministic upper bounds on the number of saved tagged candidates and the size of the audit class. Freeze a nonempty finite class of audit functions $\mathcal{H} = \{h_1, \ldots, h_d\}$ with $|h_j| \leq 1$, padding its vector representation with zero functions when necessary to reach dimension d. It contains the union of the finite preservation and repair test functions at all saved path points. For example, it contains the generated functions $s_\theta g\phi(a_\theta)$ and $d_\theta g\mathbf{1}\{q_i \in I\}$. All bin boundaries and all functions of X in this union are fixed before the following samples are observed. Predictor dependence is therefore permitted: it is frozen dependence on construction-data predictors, not adaptation to audit labels.

Use three fresh, mutually independent audit-side samples: $U_G \sim (P^X_{\mathrm{tgt}})^{m_G}$, $U_P \sim (P^X_{\mathrm{tgt}})^m$, and $S_A \sim P^{n_A}_{\mathrm{src}}$. All three are independent of construction. No target labels are observed. The same S_A is used for the quadratic audit and source screen; simultaneous events, rather than independence of those two uses, justify the reuse.

Set $h(x) = (h_1(x), \ldots, h_d(x))^\top$. For $0 < \delta \leq 1/16$, use the regularization level $\lambda = 32d\log(8d/\delta)/m_G$ and estimate the target Gram matrix by

$$\widehat{G} = \frac{1}{m_G} \sum_{x \in U_G} h(x)h(x)^\top.$$

The regularized matrix is $B = \widehat{G} + \lambda I_d$, and the normalized features are $f(x) = B^{-1/2}h(x)$. Let $c_B = \max\{1, \max_j B_{jj}\} \leq 1 + \lambda$. All these quantities are observable, including the predetermined coordinate bound d. Linear redundancy is allowed. A smaller basis can be used if the coefficient conversion for all original tests is carried through; one cannot silently replace d by an empirical rank.

Quadratic certificate and audits. The following population quantities are used for analysis and are not supplied to the selector. For a tagged candidate $i = (\theta, t)$, abbreviate $a = a_\theta$, $s = s_\theta$ and $q = q_i$. Define

$$A_\theta = \|\mathbb{E}_{P_{\mathrm{src}}}[wsf(Y-a)]\|_2^2, \tag{E.2}$$

$$P_i = \mathbb{E}_{P_{\mathrm{tgt}}}[s(q-a)^2],$$

$$R_i = \|\mathbb{E}_{P_{\mathrm{src}}}[wd_\theta f(Y-q)]\|_2^2, \tag{E.3}$$

$$J_i = A_\theta + P_i + R_i.$$

The anchor and repair terms are quadratic energies of residual directions, after target geometry normalization. The preservation term is exactly the population counterpart of the original target preservation potential. The all-repair convention makes $A_\theta = P_i = 0$. The common calibration criterion is

$$C_{\mathcal{H}}(q) = \max_{h\in\mathcal{H}} |\mathbb{E}_{P_{\mathrm{tgt}}}[h(X)(Y - q(X))]|.$$

The original finite $C_{\mathrm{tgt}}(q_i;\theta)$ is at most $C_{\mathcal{H}}(q_i)$ by the union construction.

Put $L = 2\lceil \log_2(8N/\delta)\rceil + 1$, $b = \lfloor n_A/L \rfloor \geq 2$, $\rho = \frac{2W\sqrt{d}}{b}$, and $t_0 = \log(32N/\delta)$. Divide S_A into L disjoint blocks of size b, discarding the remainder for the quadratic statistic only. For each tagged candidate define the $2d$-dimensional residual vector

$$V_i(x,y) = w(x)\begin{pmatrix} s_\theta(x)f(x)(y - a_\theta(x)) \\ d_\theta(x)f(x)(y - q_i(x)) \end{pmatrix}.$$

Write $Z_r = (X_r, Y_r)$ for the source audit observations. For block ℓ compute

$$T_{i,\ell} = \frac{\|\sum_{r\in\ell} V_i(Z_r)\|_2^2 - \sum_{r\in\ell}\|V_i(Z_r)\|_2^2}{b(b-1)}, \qquad T_i = \mathrm{median}_\ell\, T_{i,\ell}.$$

This block-median construction is related to median-of-means robustification for U-statistics (Joly & Lugosi, 2016). Each block statistic equals the sum of pairwise inner products over distinct observations divided by $b(b-1)$, so no repeated input values or explicit quadratic pair enumeration are required. It estimates $A_\theta + R_i$, not the noisy empirical squared norm of a sample mean.

On U_P compute $\widehat{P}_i = m^{-1}\sum_{x\in U_P} s_\theta(x)(q_i(x) - a_\theta(x))^2$. Set

$$U_i = \min\left\{3, 2[T_i]_+ + 44\rho + 2\widehat{P}_i + \frac{4t_0}{m}\right\}. \tag{E.4}$$

There is no unknown residual, variance, density discrepancy, or optimal partition parameter in this formula. The constants are conservative proof constants, not an assertion of a sharp practical threshold.

Remark E.1 (Key technical device). The main technical step behind the quadratic certificate is an order-two U-statistic combined with a block-median step, analogous to median-of-means robustification. U-statistics are classical tools for unbiased estimation of quadratic functionals; here, the statistic is used not merely for estimation, but to construct a common certificate for joint selection over candidate preserve-or-repair partitions and their path predictors. Using only off-diagonal pairs removes the diagonal self-product bias of the plug-in squared norm of the empirical residual mean. The resulting block-median statistic admits, uniformly over all saved candidates, an additive comparison slack of order $W\sqrt{d}/b$, enabling a single oracle comparison across both candidate partitions and predictors.

Source screen and selection rule. For each tagged candidate, let $q_0 = q_\theta^{(0)}$. Using all n_A observations in S_A, compute the mean Brier difference $\widehat{\Delta}_i = n_A^{-1}\sum_{(X,Y)\in S_A}\{(Y - q_i(X))^2 - (Y - q_0(X))^2\}$ and the mean squared prediction change $\widehat{D}_i = n_A^{-1}\sum_{(X,Y)\in S_A}(q_i(X) - q_0(X))^2$. Set $x_A = t_0/n_A$ and $D_i^+ = \min\{1, 2\widehat{D}_i + 4x_A\}$. The simultaneous radius is $r_i = \sqrt{8D_i^+ x_A} + 2x_A$. Retain i if $\widehat{\Delta}_i + r_i \leq 0$, and retain every initializer unconditionally. Denote the resulting nonempty tagged family by $\mathcal{E}$. This is the same paired source-Brier nonincrease requirement as in the original

method, with a fully specified simultaneous radius that localizes with the size of the update. It is not a target-Brier safety condition.

Select the eligible candidate with the smallest upper bound:

$$\hat{i} = (\widehat{\theta}_{\mathrm{str}}, \hat{t}) \in \arg\min_{i \in \mathcal{E}} U_i.$$

The returned predictor is $\widehat{q}_{\mathrm{str}} = q_{\hat{i}}$, with bound $\widehat{U} = U_{\hat{i}}$ and calibration certificate $\widehat{C}_{\mathrm{str}} = \min\{1, \sqrt{3c_B\widehat{U}}\}$. Joint minimization is part of this auxiliary construction. Coverage may be used only to break exact ties.

E.2 Certificate for Joint Partition and Predictor Selection (Lem. E.2 and Proof)

Lemma E.2 (Quadratic certificate for joint partition and predictor selection)**.** *Under* (E.1) *and the stated sample protocol, conditionally on arbitrary preprocessing and construction data, with probability at least* $1 - \delta$ *all the following hold:*

1. *The target audit geometry satisfies* $\mathbb{E}_{P_{\mathrm{tgt}}}[f(X)f(X)^\top] \preceq 2I_d$.
2. *Every saved candidate satisfies*

$$J_i \le U_i \le 3J_i + 88\rho + 8t_0/m, \qquad C_{\mathcal{H}}(q_i)^2 \le 3c_B J_i.$$

3. *Every eligible candidate is source-Brier safe relative to its own initializer. In particular,*

$$\mathsf{Risk}_{P_{\mathrm{src}}}(\widehat{q}_{\mathrm{str}}) \le \mathsf{Risk}_{P_{\mathrm{src}}}(q^{(0)}_{\widehat{\theta}_{\mathrm{str}}}), \qquad C_{\mathrm{tgt}}(\widehat{q}_{\mathrm{str}}; \widehat{\theta}_{\mathrm{str}}) \le C_{\mathcal{H}}(\widehat{q}_{\mathrm{str}}) \le \widehat{C}_{\mathrm{str}}.$$

4. *Selection competes across* all *eligible candidates* θ *and their path predictors:*

$$J_{\hat{i}} \le \widehat{U} \le 3 \min_{i \in \mathcal{E}} J_i + \frac{176W\sqrt{d}}{b} + \frac{8t_0}{m}.$$

No anchor-correctness or residual-realizability assumption is needed for these conclusions. The comparator family $\mathcal{E}$ is random; the simultaneous event justifies the comparison. The lemma does not by itself assert that the constructed paths contain a good candidate; Lem. E.9 below proves a sufficient structural case.

The proof proceeds through four lemmas controlling the target geometry, the quadratic envelope, the block-median statistic, and the source screen.

Lemma E.3. *Let* $G = \mathbb{E}_{P_{\mathrm{tgt}}}[hh^\top]$. *With probability at least* $1 - \delta/4$ *over* U_G, $G \preceq 2B$. *Consequently, for every square-integrable scalar function* v,

$$\|\mathbb{E}_{P_{\mathrm{tgt}}}[fv]\|_2^2 \le 2\mathbb{E}_{P_{\mathrm{tgt}}}[v^2]. \tag{E.5}$$

Proof. If $\lambda > d$, then $G \preceq dI \preceq B$ deterministically. Otherwise set $A = G + \lambda I$, $u = A^{-1/2}h$, $C = \mathbb{E}[uu^\top] \preceq I$, and $\kappa = d/\lambda \ge 1$. Then $\|u\|^2 \le \kappa$. For $Z = uu^\top - C$, $\|Z\|_{\mathrm{op}} \le \kappa$ and $\mathbb{E}[Z^2] = \mathbb{E}[(uu^\top)^2] - C^2 \preceq \kappa C \preceq \kappa I$. Applying the matrix Bernstein inequality (Tropp, 2012, Thm. 6.1(ii)) to both signs gives

$$\mathbb{P}\{\|m_G^{-1}\sum Z_j\|_{\mathrm{op}} > 1/2\} \le 2d\exp\{-3m_G/(28\kappa)\} \le \delta/4.$$

On the complement, $\widehat{G} + \lambda I \succeq \frac{1}{2}(G + \lambda I)$, hence $G \preceq 2B$. For (E.5), dualize the Euclidean norm and use Cauchy–Schwarz: $|\mathbb{E}[(v_0^\top f)v]|^2 \le \mathbb{E}[(v_0^\top f)^2]\mathbb{E}[v^2] \le 2\mathbb{E}[v^2]$ for every unit vector v_0. □

Lemma E.4. *On* $G \preceq 2B$, $0 \le J_i \le 3$ *and* $C_{\mathcal{H}}(q_i)^2 \le 3c_B J_i$.

Proof. By exact transport, the two source means in (E.2) and (E.3) are target means. Since s and d_θ have disjoint supports, (E.5) implies

$$A_\theta + R_i \le 2\mathbb{E}_{P_{\mathrm{tgt}}}[s(\eta - a)^2 + d_\theta(\eta - q)^2] \le 2, \qquad P_i \le 1.$$

For an original test h_j, let e_j be the jth coordinate vector and decompose its residual into anchor, preservation, and repair parts. Because $h_j = e_j^\top B^{1/2} f$,

$$\begin{aligned}|\mathbb{E}_{P_{\mathrm{tgt}}}[h_j(Y-q)]| &\le \sqrt{B_{jj}}(\sqrt{A_\theta}+\sqrt{R_i}) + |\mathbb{E}_{P_{\mathrm{tgt}}}[s h_j(a-q)]| \\ &\le \sqrt{B_{jj}}(\sqrt{A_\theta}+\sqrt{R_i}) + \sqrt{P_i}.\end{aligned}$$

Squaring, using $(x+y+z)^2 \le 3(x^2+y^2+z^2)$, and maximizing over j proves the result. □

Lemma E.5. *On the geometry event, condition on U_G and all construction data. For every fixed saved candidate put $r = A_\theta + R_i$. Its block statistic satisfies*

$$\mathbb{E}[T_{i,\ell}] = r, \qquad \mathrm{Var}(T_{i,\ell}) \le \frac{8Wr}{b} + \frac{8W^2 d}{b(b-1)}.$$

With conditional probability at least $1-\delta/8$, simultaneously for all saved candidates, $r \le 2[T_i]_+ + 44\rho \le 3r + 88\rho$.

Proof. Let $V = V_i(X,Y)$, $\mu = \mathbb{E}_{P_{\mathrm{src}}}[V]$, and $M = \mathbb{E}_{P_{\mathrm{src}}}[VV^\top]$. Write $\Gamma_s = \mathbb{E}_{P_{\mathrm{tgt}}}[s f f^\top]$, $\Gamma_d = \mathbb{E}_{P_{\mathrm{tgt}}}[d_\theta f f^\top]$. Disjoint supports give zero off-diagonal blocks, and $w^2 \le Ww$ gives $M \preceq W\,\mathrm{diag}(\Gamma_s, \Gamma_d)$, $\|M\|_{\mathrm{op}} \le 2W$, and

$$\mathrm{tr}(M^2) \le W^2\{\mathrm{tr}\,\Gamma_s^2 + \mathrm{tr}\,\Gamma_d^2\} \le W^2\,\mathrm{tr}(\Gamma_s+\Gamma_d)^2 \le 4W^2 d.$$

For the kernel $H(Z,Z') = V(Z)^\top V(Z')$, its mean is $\|\mu\|^2 = r$, and its first projection is $V^\top\mu$. Moreover, $\mathbb{E}[(V^\top\mu)^2] \le 2Wr$ and $\mathbb{E}[H^2] = \mathrm{tr}(M^2) \le 4W^2 d$. The standard variance formula for an order-two U-statistic (Serfling, 1980, Chapter 5) therefore gives the stated bound. As $b \ge 2$, it is at most $8Wr/b + 16W^2 d/b^2$. Put $u = r/\rho$. Chebyshev implies

$$\mathbb{P}\{|T_{i,\ell} - r| > r/2 + 22\rho\} \le \frac{4(u+1)}{(u/2+22)^2} \le 1/8.$$

The last inequality follows by subtracting $32(u+1)$ from the denominator and completing the square. If the median fails, at least half of the independent blocks fail. Its failure probability is at most $2^{-L/2} \le \delta/(8N)$. A union bound proves the simultaneous statement. The median interval implies both bounds by inversion and by $[T_i]_+ \le 3r/2 + 22\rho$. □

Let $r_{\mathrm{src}}(q) = \mathbb{E}_{P_{\mathrm{src}}}[(q-\eta)^2]$ denote source excess Brier risk.

Lemma E.6 (Source safety and the retained initializer)**.** *With probability at least $1-\delta/8$ over S_A, simultaneously for all saved candidates, $|\widehat{\Delta}_i - \Delta_i| \le r_i$, where $\Delta_i = \mathsf{Risk}_{P_{\mathrm{src}}}(q_i) - \mathsf{Risk}_{P_{\mathrm{src}}}(q_\theta^{(0)})$. Thus every eligible candidate is source safe. Moreover, for every $i = (\theta, t)$ there is a candidate j on the* same *path that remains eligible and satisfies*

$$r_{\mathrm{src}}(q_j) \le 3 r_{\mathrm{src}}(q_i) + 300 t_0/n_A. \tag{E.6}$$

One may take $q_j = q_i$ when retained and $q_j = q_\theta^{(0)}$ otherwise. No population improvement margin is an input or assumption.

Proof. Put $D = \mathbb{E}_{P_{\mathrm{src}}}[(q_i - q_0)^2]$ and $x = t_0/n_A$. For any $Z \in [0,1]$, the scalar Bernstein inequality (Boucheron et al., 2013, Sec. 2.8), together with $\mathrm{Var}(Z) \le \mathbb{E}[Z]$, implies, except with probability $2e^{-t_0}$, $\mathbb{E}[Z] \le 2\overline{Z} + 4x$ and $\overline{Z} \le \frac{3}{2}\mathbb{E}[Z] + \frac{5}{3}x$. Apply this to $(q_i - q_0)^2$ to obtain $D \le D^+ \le 3D + 8x$. The paired loss difference is in $[-1,1]$ and has variance at most $4D$. Thus, except with probability $2e^{-t_0}$, $|\widehat{\Delta} - \Delta| \le \sqrt{8Dx} + \frac{4}{3}x \le r_i$. Also $r_i \le 5\sqrt{Dx} + 10x \le 8\sqrt{(r+r_0)x} + 10x$, where $r = r_{\mathrm{src}}(q_i)$, $r_0 = r_{\mathrm{src}}(q_0)$, using $D \le 2(r+r_0)$. If rejected, $\widehat{\Delta} + r_i > 0$, hence $\Delta + 2r_i > 0$ and

$$r_0 < r + 16\sqrt{(r+r_0)x} + 20x \le r + \tfrac{1}{2}(r+r_0) + 148x.$$

It follows that $r_0 \le 3r + 296x$, proving (E.6). A union bound over the two concentration events and all N candidates costs $4Ne^{-t_0} = \delta/8$. Source safety follows from the upper confidence bound defining the source-Brier screen; initializers are equal to their own baselines. □

Proof of Lem. E.2. By the same scalar Bernstein argument, except with probability $2Ne^{-t_0} \leq \delta/16$, simultaneously $P_i \leq 2\widehat{P}_i + 4t_0/m \leq 3P_i + 8t_0/m$. Combine this with Lems. E.3–E.6. The sum of the failure budgets is less than δ. Since $J_i \leq 3$, truncating U_i at three preserves coverage. The upper oracle inequality is unchanged by truncation. Minimizing U_i over the random nonempty family $\mathcal{E}$ gives

$$J_{\hat{i}} \leq U_{\hat{i}} \leq \min_{i \in \mathcal{E}} U_i \leq 3 \min_{i \in \mathcal{E}} J_i + 88\rho + 8t_0/m.$$

Lem. E.4 gives the common calibration certificate. Since the common class contains the original audit functions, this certificate also bounds the original candidate's calibration error over those functions. □

E.3 Adaptation to an Unknown Repair Region (Lem. E.9 and Proof)

The key statistical question is whether selective inheritance can adapt to a small repair region without knowing that region in advance. The result below shows that it can: the estimation term scales with the number of cells that actually require repair, while the procedure simultaneously searches over candidate partitions and predictors.

The general lemma allows arbitrary construction paths. To obtain an explicit rate rather than a path-oracle statement, one must also prove that a useful candidate is generated. The following specialization does so without replacing the projected residual-update formula by a regression oracle. The residual-update argument is the finite-cell analogue of the squared-loss boosting perspective on multicalibration (Globus-Harris et al., 2023); low-degree multicalibration provides a related moment-based formulation (Gopalan et al., 2022a).

For the adaptation analysis, abbreviate the repair-cell count k_{rep} as k. Fix a preprocessing-measurable candidate θ_*, which is not supplied to the selector. Write $s_* = s_{\theta_*}$, $a_* = a_{\theta_*}$, and $D_* = \{s_* = 0\}$. Suppose:

1. **Anchor validity in the audited directions.** For every test function that can appear in the fixed common class of audit functions,
$$\mathbb{E}_{P_{\mathrm{tgt}}}[s_* h(X)(Y - a_*(X))] = 0. \tag{E.7}$$
This condition holds almost surely over construction. More explicitly, it suffices that all classes of audit functions determined by construction data are subsets of a fixed master class for which (E.7) holds. Pointwise equality $\eta = a_*$ on the stable region is sufficient but *not necessary*. The master class is not charged as a statistical search class: once construction is conditioned on, only its actual finite subclass is audited.
2. **A finite repair part.** D_* is partitioned into k preprocessing-fixed disjoint cells $C_1, \ldots, C_k$ of positive source probability. On each cell, $\eta(x) = \eta_j$ and $w(x) = w_j > 0$ are constant. There is no uniform lower bound on w_j or on the cell probabilities. On the stable region, X, the scores, and the anchor can be continuous and arbitrarily complex.
3. **Initializer and repair directions.** The candidate is initialized at a_* on the stable region and at $1/2$ on D_*. Cell indicators are included as repair directions. If the original class of repair functions uses current-score bins, these directions are available whenever the cell is a base group: a cell-constant predictor belongs to one bin on the whole cell. Extra supported directions on this path may be omitted in this specialization.
4. **A saved sufficiently converged endpoint.** With $S_0 \sim P_{\mathrm{src}}^{n_0}$ used for construction, apply the same repair update (4.1) in the cell with largest absolute residual correlation until, on every observed repair cell, $|q_j - \overline{Y}_j| \leq \varepsilon$, and save this endpoint as well as the initializer. Unobserved repair cells remain at $1/2$. This is an observable stopping rule with a prescribed $\varepsilon > 0$. It is not a claim about a fixed 50-round cap.

The anchor equality in (E.7) describes a favorable family for adaptation. Neither these equalities nor the identity of θ_* nor k is an input to the algorithm. Certificate validity in Lem. E.2 does not require them.

The whole candidate family may contain additional arbitrary A-PAIR paths. The deterministic bounds N and d are used in all tuning formulas, so the bounds remain valid when expectations also include construction. Saving a prescribed number of snapshots, rather than every numerical iteration, makes N finite even when the number of internal updates depends on the data.

Lemma E.7 (Finite completion and local estimation cost)**.** *The repair process in the fourth condition terminates after finitely many updates almost surely. If q_* is its saved endpoint and $r_{D_*}(q) = \mathbb{E}_{P_{\mathrm{src}}}[\mathbf{1}_{D_*}(q-\eta)^2]$, then*

$$\mathbb{E}[r_{D_*}(q_*)] \le \frac{k}{n_0+1} + 2\varepsilon^2. \tag{E.8}$$

The endpoint and the initializer both equal a_* on the stable region.

Proof. For each observed cell put $\widehat{p}_j = \frac{1}{n_0}\sum_{(X,Y)\in S_0} w(X)\mathbf{1}_{C_j}(X) > 0$. Since w is constant on each cell, its weighted label average equals the ordinary average $\overline{Y}_j$. The selected correlation is $c_j = \widehat{p}_j(\overline{Y}_j - q_j)$, so the same repair formula gives

$$q_j^+ = q_j + \frac{\widehat{p}_j}{W}(\overline{Y}_j - q_j).$$

Because $0 < \widehat{p}_j/W \le 1$, this is a convex combination and clipping is inactive. It preserves cell-constant predictions and changes no other cell. Preservation correlations vanish identically throughout this path.

Let $p_{\min}$ be the smallest positive $\widehat{p}_j$. For the observable optimization gap $F(q) = \sum_{j:\widehat{p}_j>0}\widehat{p}_j(q_j - \overline{Y}_j)^2$, the maximum-correlation choice satisfies

$$\max_j c_j^2 \ge \frac{p_{\min}}{k}F(q), \qquad F(q) - F(q^+) \ge c_j^2/W.$$

Thus F contracts geometrically by a factor at most $1 - p_{\min}/(kW)$. The condition $F \le p_{\min}\varepsilon^2$ implies the stopping rule. If there are no observed repair cells, the rule is satisfied without any update. This proves finite completion; $p_{\min}$ enters this computational analysis but is not a tuning parameter or a statistical remainder.

Define $\bar{q}$ to equal the cell mean on an observed repair cell, $1/2$ on an unobserved repair cell, and a_* on the stable region. If N_j is the source count in cell j, then, conditionally on $N_j = m$, $\mathbb{E}[(\bar{q}_j - \eta_j)^2 \mid N_j = m] \le \frac{1}{2(m+1)}$. For $m \ge 1$ this follows from $\eta_j(1-\eta_j)/m \le 1/[2(m+1)]$; for $m = 0$ it follows from $(1/2 - \eta_j)^2 \le 1/4$. For a binomial count with cell probability p_j, expanding the probabilities and using $\binom{n_0}{r}/(r+1) = \binom{n_0+1}{r+1}/(n_0+1)$ gives $p_j\mathbb{E}[(N_j+1)^{-1}] = \frac{1-(1-p_j)^{n_0+1}}{n_0+1} \le \frac{1}{n_0+1}$, which implies $\mathbb{E}[r_{D_*}(\bar{q})] \le k/[2(n_0+1)]$. The saved endpoint differs from $\bar{q}$ by at most ε on observed cells and agrees elsewhere. Apply $(x+y)^2 \le 2x^2 + 2y^2$ and sum source cell masses to prove (E.8). □

Remark E.8 (The source screen localizes to the changed region). If $q_i = q_0$ off a fixed region D, the proof of Lem. E.6 also gives

$$r_D(q_j) \le 3r_D(q_i) + 300t_0/n_A, \qquad q_j \in \{q_i, q_0\} \text{ eligible}, \tag{E.9}$$

where $r_D(q) = \mathbb{E}_{P_{\mathrm{src}}}[\mathbf{1}_D(q-\eta)^2]$. Indeed, the population paired loss difference equals $r_D(q_i) - r_D(q_0)$, and $\mathbb{E}_{P_{\mathrm{src}}}[(q_i - q_0)^2] \le 2\{r_D(q_i) + r_D(q_0)\}$. All other steps are unchanged. This localization is essential: it prevents unmodelled pointwise anchor error on the preserved region from entering the adaptation rate. Only its audited moments need be zero.

Lemma E.9 (Adaptation to an unknown repair region with retained source safety)**.** *Under the conditions above, the audited selector retains all the validity and safety conclusions of Lem. E.2, including for distributions outside the favorable family. On the favorable family described by* (E.7) *and the k repair-cell conditions,*

$$\mathbb{E}[\widehat{U}] \le \min\Big\{3, \frac{18Wk}{n_0+1} + 36W\varepsilon^2 + \frac{1800Wt_0}{n_A} + \frac{176W\sqrt{d}}{b} + \frac{8t_0}{m} + 3\delta\Big\}. \tag{E.10}$$

All expectations include construction, geometry, preservation, and source monitoring samples. No improvement-margin assumption for passing the source-Brier screen is needed.

Proof. On the simultaneous good event from Lem. E.2, apply (E.9) to the saved endpoint q_*. There is an eligible $q_\dagger \in \{q_*, q_{\theta_*}^{(0)}\}$ with $r_{D_*}(q_\dagger) \le 3r_{D_*}(q_*) + 300t_0/n_A$. Both predictors equal a_* on

the stable region. Thus their preservation term is zero, and (E.7) implies $A_{\theta_*} = 0$ even though the whitening matrix depends on the independent geometry sample. By (E.5) and exact transport,

$$J_\dagger = \|\mathbb{E}_{P_{\mathrm{tgt}}}[\mathbf{1}_{D_*} f(\eta - q_\dagger)]\|^2 \le 2\mathbb{E}_{P_{\mathrm{tgt}}}[\mathbf{1}_{D_*}(\eta - q_\dagger)^2] \le 2W r_{D_*}(q_\dagger).$$

Consequently, the joint selector gives

$$\widehat{U} \le 18W r_{D_*}(q_*) + \frac{1800W t_0}{n_A} + \frac{176W\sqrt{d}}{b} + \frac{8t_0}{m}.$$

Use Lem. E.7. On the bad event use the deterministic bound $\widehat{U} \le 3$ and its probability at most δ. This proves (E.10). □

E.4 General Rate for a Finite Class of Audit Functions (Lem. E.10 and Proof)

The same auxiliary construction also yields a calibration-certificate bound on a general input space. Let $n = n_0 + n_A$, take comparable construction and source-audit split sizes, and write $\ell := \log\{32Nd(n+m+m_G+1)^2/\delta\}$. Here N and d are the deterministic bounds in App. E.1; m and m_G are the preservation-audit and geometry sample sizes.

Lemma E.10 (Adaptation to an unknown repair region with a finite audit class)**.** *Assume exact covariate transport with known ratio $w = dP^X_{\mathrm{tgt}}/dP^X_{\mathrm{src}} \in [0, W]$, $0 < \delta \le 1/16$, and the independent preprocessing and audit protocol of App. E.1. For sufficiently large universal constants in $n \gtrsim \ell$ and $m_G \gtrsim d\ell$, the audited preserve-or-repair selector returns $(\widehat{q}_{\mathrm{str}}, \widehat{\theta}_{\mathrm{str}})$ and a computable $\widehat{C}_{\mathrm{str}}$ such that, with probability at least $1 - \delta$,*

$$C_{\mathcal{H}}(\widehat{q}_{\mathrm{str}}) \le \widehat{C}_{\mathrm{str}}, \qquad \mathsf{Risk}_{P_{\mathrm{src}}}(\widehat{q}_{\mathrm{str}}) \le \mathsf{Risk}_{P_{\mathrm{src}}}(q^{(0)}_{\widehat{\theta}_{\mathrm{str}}}).$$

These conclusions do not require a correct anchor.

For the adaptation statement, suppose a preprocessing-fixed candidate θ_*, unknown to the selector, has zero anchor-residual moments on its preserve region for every direction that can enter the common class of audit functions. Suppose its repair region consists of k_{rep} fixed cells, on each of which η and the positive density ratio are constant. On this path, initialize at the anchor on the preserve region and at $1/2$ on the repair region, use cell-indicator repair directions, and save an endpoint within $n^{-1/2}$ of the empirical label mean on every observed cell; unobserved cells remain at $1/2$. Under the conservative confidence allocation in App. E.2, a universal constant C satisfies

$$\mathbb{E}[\widehat{C}^2_{\mathrm{str}}] \le C \min\left\{1, \frac{W k_{\mathrm{rep}}}{n} + \frac{W\sqrt{d}\,\ell}{n} + \frac{\ell}{m}\right\}.$$

Proof of Lem. E.10. The certificate and source-safety conclusions follow from Lem. E.2. To avoid a constant failure-probability term in an expected-rate display, run all tuning formulas at $\delta' = \frac{\delta}{(n_0+n_A+m+m_G+1)^2}$. Let $n = n_0 + n_A$, take n_0, n_A comparable to n, choose $\varepsilon^2 \le 1/n$, and assume $n_A \ge 2L$ and $m_G \ge 32d\log(8d/\delta')$. Then $\lambda \le 1$, $c_B \le 2$, and for a universal constant C,

$$\mathbb{E}[\widehat{C}^2_{\mathrm{str}}] \le C \min\left\{1, \frac{Wk}{n} + \frac{W\sqrt{d}\log(8N/\delta')}{n} + \frac{\log(32N/\delta')}{m}\right\}.$$

The source-screen logarithmic term is absorbed because $\sqrt{d} \ge 1$. The trivial upper bound of one may always be imposed on the whole display. The extra m_G unlabeled observations are explicitly part of the experiment; they are not free knowledge of target covariance. In rate notation this is $\widetilde{O}\{W(k+\sqrt{d})/n + 1/m\}$ for the *squared* calibration certificate. The certificate for unsquared residual moments has the square root of this rate. The excess-Brier comparison in App. E.5 additionally uses complete finite-cell geometry.

The logarithms above are bounded by the lemma's ℓ up to universal constants; the assumed lower bounds on n and m_G ensure the stated sample-size conditions. Since $\delta' \le \delta$, the confidence statement remains at least $1 - \delta$. This proves the stated rate for the finite audit class. □

E.5 Statistical Separation from Full Refitting (Thm. E.11 and Proof)

Let X be uniform on $[K]$, with $Y \mid X = j \sim \mathrm{Bernoulli}(\eta_j)$, and let source and target coincide. Assume the state is determined by the observed score pair. The covariate law is known to both procedures. Fix an anchor $a \in [1/4, 3/4]^K$ and a finite, label-independent collection $\mathcal{D}$ of repair regions. Let $\mathcal{F}_k$ contain all $\eta \in [1/4, 3/4]^K$ for which some $D \in \mathcal{D}$ satisfies $|D| \le k$ and $\eta_j = a_j$ for every $j \notin D$. We measure excess Brier risk by $\mathcal{E}_\eta(q) = K^{-1} \sum_{j=1}^K (q_j - \eta_j)^2$ and assume $\mathcal{F}_k$ is nonempty. The candidate regions may be discrepancy-threshold regions; no correct region is provided to the selector.

Specialized audited procedure. Use the cell indicators as the common basis $h_j(x) = \mathbf{1}\{x = j\}$. Its known Gram matrix is $G = K^{-1} I_K$, so take the normalization matrix $B = G$ and $f(x) = \sqrt{K}\, e_x$, where e_x is the xth coordinate vector, without estimating geometry. All finite-state audit functions are linear combinations of this basis. On each candidate path initialize at a on the preserve region and $1/2$ on the repair region, and use only cell-indicator repair updates. Save the initializer and an endpoint whose distance to the empirical label mean is at most ε on observed repair cells; unobserved cells remain at $1/2$. Only finitely many snapshots are saved per candidate, irrespective of the number of internal updates. Let N bound their total number deterministically.

Split the n source labels into comparable construction and audit sizes n_0, n_A. Use the localized paired source-Brier screen in Rem. E.8, and jointly minimize its quadratic envelope upper bound. Since the covariate law and the statewise predictions are known, preservation averages can be evaluated exactly. In fact they are identically zero on the specified cell-indicator paths. Use $\varepsilon^2 \le 1/n$ and internal failure level $\delta' = \delta/(n+1)^2$, $0 < \delta \le 1/16$. Set $d = K$ in the quadratic source audit and use δ' in its tuning rules and source screen. Replace Eq. (E.4) by the exact-preservation score $U_i = \min\{3,\, 2[T_i]_+ + 44\rho + P_i\}$, with $P_i = 0$. Thus neither an estimated geometry nor a target-sampling term enters this specialization.

Theorem E.11 (Statistical separation from full refitting; formal version of Thm. 5.1)**.** *For the specialized audited procedure above, there are universal $c, C > 0$ such that, for $n \ge \max\{K, c\ell\}$ with $\ell = \log\{32N(n+1)^2/\delta\}$,*

$$\sup_{\eta \in \mathcal{F}_k} \mathbb{E}[\mathcal{E}_\eta(\widehat{q}_{\mathrm{str}})] \le C \min\left\{1, \frac{k + \sqrt{K}\,\ell}{n}\right\}.$$

Define the comparator $\widehat{q}_{\mathrm{full}}$ using all n labels: it is the empirical positive-label frequency on every observed state and $1/2$ on every unobserved state. Then, pointwise for every $\eta \in [1/4, 3/4]^K$,

$$\frac{K}{48n} \le \mathbb{E}[\mathcal{E}_\eta(\widehat{q}_{\mathrm{full}})] \le \frac{K}{2(n+1)}.$$

In particular, if $k + \sqrt{K}\,\ell = o(K)$ along a sequence of these models, then

$$\sup_{\eta \in \mathcal{F}_k} \frac{\mathbb{E}[\mathcal{E}_\eta(\widehat{q}_{\mathrm{str}})]}{\mathbb{E}[\mathcal{E}_\eta(\widehat{q}_{\mathrm{full}})]} \longrightarrow 0.$$

Thm. E.11 identifies a statistical regime in which selective inheritance changes the effective complexity of calibration. When only k of K cells require repair, its expected excess Brier risk is governed by the residual complexity k, together with the cost of discovering and auditing that structure, rather than by the full K-cell estimation problem. Consequently, when $k + \sqrt{K}\,\ell = o(K)$, selective inheritance achieves a vanishing risk ratio relative to full cellwise refitting under the same label budget. This establishes that selective inheritance can yield a statistical efficiency gain by converting global re-estimation into residual learning.

Proof. First connect the quadratic envelope to the comparison loss. With the known complete-cell geometry, a tagged candidate (D, q) has anchor term $A_D = K^{-1} \sum_{j \notin D} (\eta_j - a_j)^2$, preservation term $P_D(q) = K^{-1} \sum_{j \notin D} (q_j - a_j)^2$, and repair term $R_D(q) = K^{-1} \sum_{j \in D} (q_j - \eta_j)^2$. Consequently, for every candidate and every η, $\mathcal{E}_\eta(q) \le 2\{A_D + P_D(q) + R_D(q)\} = 2J_D(q)$. The exact geometry satisfies $\mathbb{E}[f f^\top] = I_K$, so the covariance inequalities in the quadratic-audit proof

hold without a geometry-failure event. The source-screen and local-construction arguments are unchanged, with $W = 1$ and $d = K$. An appropriate candidate in $\mathcal{F}_k$ has $A_D = 0$; its saved endpoint and initializer both have $P_D = 0$. The proof of Lem. E.9, with exact preservation averages, therefore gives

$$\mathbb{E}[\widehat{U}] \le C_0 \left\{ \frac{k}{n_0 + 1} + \varepsilon^2 + \frac{\log(32N/\delta')}{n_A} + \frac{\sqrt{K}\log(8N/\delta')}{n_A} \right\} + 3\delta'.$$

On the simultaneous audit event, $\mathcal{E}_\eta(\widehat{q}_{\mathrm{str}}) \le 2\widehat{U}$; off this event, $\mathcal{E}_\eta(\widehat{q}_{\mathrm{str}}) \le 1$. Thus $\mathbb{E}[\mathcal{E}_\eta(\widehat{q}_{\mathrm{str}})] \le 2\mathbb{E}[\widehat{U}] + \delta'$. Comparable splits and $\varepsilon^2 \le 1/n$ prove the first claim.

For the comparator let N_j be the state-j count. Conditional on $N_j = r > 0$, its squared estimation error has expectation $\eta_j(1-\eta_j)/r$. For $r = 0$, its error is $(1/2 - \eta_j)^2$. Since $\eta_j(1-\eta_j) \ge 3/16$,

$$\mathbb{E}[(\widehat{q}_{\mathrm{full},j} - \eta_j)^2] \ge \frac{3}{16}\mathbb{E}\left[\frac{\mathbf{1}\{N_j > 0\}}{N_j}\right] \ge \frac{3}{16}\mathbb{E}\left[\frac{N_j}{(N_j+1)^2}\right].$$

For $M \sim \mathrm{Binomial}(n, p)$ and $B \sim \mathrm{Binomial}(n-1, p)$, expanding the binomial probabilities and using $r\binom{n}{r} = n\binom{n-1}{r-1}$ gives $\mathbb{E}[Mg(M)] = np\,\mathbb{E}[g(B+1)]$. Write $\mu = n/K \ge 1$. Applying this identity with $g(r) = (r+1)^{-2}$ and $B_j \sim \mathrm{Binomial}(n-1, 1/K)$ yields

$$\mathbb{E}\left[\frac{N_j}{(N_j+1)^2}\right] = \mu\mathbb{E}\left[\frac{1}{(B_j+2)^2}\right] \ge \frac{\mu}{(\mu+2)^2} \ge \frac{1}{9\mu}.$$

The middle inequality follows from Jensen's inequality for the convex function $x \mapsto (x+2)^{-2}$, together with $\mathbb{E}[B_j] \le \mu$. It follows that each state has expected squared error at least $K/(48n)$, and averaging the K states preserves this lower bound.

For the upper bound, conditionally on $N_j = r$, the expected squared error is at most $1/[2(r+1)]$, including $r = 0$. Multiplying $\binom{n}{r}/(r+1) = \binom{n+1}{r+1}/(n+1)$ by $p^r(1-p)^{n-r}$ and summing over $r = 0, \dots, n$ gives $\mathbb{E}[(N_j+1)^{-1}] = [1-(1-p)^{n+1}]/[(n+1)p]$. With $p = 1/K$, this becomes $\mathbb{E}[\frac{1}{N_j+1}] = \frac{K}{n+1}\{1-(1-1/K)^{n+1}\} \le \frac{K}{n+1}$ and gives the stated upper bound after averaging. Combining the pointwise lower bound with the uniform adaptive upper bound proves the risk-ratio claim. □

F Implementation Details

F.1 Anchor Construction and Candidate Stable Regions

The predecessor logit may retain distinctions lost after fine-tuning, but its numerical location and scale need not match those of the fine-tuned logit. Let $U_{\mathrm{const}} = \{X_i^u\}_{i=1}^{m_{\mathrm{const}}}$ denote the unlabeled target construction sample, which is not used for holdout evaluation. For each prespecified class $\mathcal{G}_r$ of increasing maps, we fit

$$\hat{\psi}_r \in \arg\min_{\psi \in \mathcal{G}_r} \frac{1}{m_{\mathrm{const}}} \sum_{i=1}^{m_{\mathrm{const}}} \{z_{\mathrm{ft}}(X_i^u) - \psi(z_{\mathrm{pre}}(X_i^u))\}^2. \tag{F.1}$$

We then define $z_{\mathrm{anc},r}(x) := \hat{\psi}_r(z_{\mathrm{pre}}(x))$ and $q_{\mathrm{anc},r}(x) := \mathrm{sigmoid}(z_{\mathrm{anc},r}(x))$. Thus, "scale-aligned" means that the transformed predecessor logit is the empirical least-squares projection of the fine-tuned logit onto the chosen increasing-map class under the target construction covariates. This alignment uses no labels and does not certify that the anchor is calibrated. We separately check whether each candidate passes the anchor audit using labeled source data.

Finite-grid anchor selection. For a finite collection of fitted anchor maps and discrepancy thresholds, Eq. (4.3) selects the candidate with the largest stable-region coverage subject to the simultaneous anchor-audit upper bound being at most γ_{anc}. If no anchor–threshold candidate is feasible, all points are assigned to repair. This validity–coverage rule makes no monotonicity assumption about the signed anchor residual as the stable region grows, and its finite-family cost enters only through the simultaneous audit.

Instantiation used in the experiments. The reported experiments use two anchor-map candidates: the identity map $\psi(z) = z$ and a positive-slope affine map fitted by Eq. (F.1). The stable region of each candidate additionally requires agreement between the predecessor and fine-tuned argmax answers. Among agreement examples, candidate stable regions include inputs with smaller aligned logit discrepancy first. For each anchor map, we construct candidates with stable-mass caps 0.05, 0.10, 0.20, 0.35, and 0.50. Among the candidates that pass the anchor audit, we select the candidate $\hat{\theta}$ with the largest coverage on the independent unlabeled target audit set. If none passes, A-PAIR uses the all-repair candidate.

F.2 Density-Ratio Estimation

For MMLU-Pro and SciQ, A-PAIR estimates the target-over-source density ratio $dP^X_{\mathrm{tgt}}/dP^X_{\mathrm{src}}$ from source-construction and unlabeled target-construction covariates. We fit an ℓ_2-regularized logistic-regression domain classifier on standardized model-score features to distinguish source ($D = S$) from target ($D = T$). Its predicted domain probability is converted to $\widehat{w}(x) = \frac{n_S}{n_T} \frac{\widehat{P}(D=T|X=x)}{1-\widehat{P}(D=T|X=x)}$. We use five-fold cross-fitting and select the inverse regularization strength from $C \in \{0.1, 1, 10\}$ by domain-classification loss. Source-construction weights use out-of-fold predictions. For the independent source audit set, the domain classifier is refit on the full construction data and applied without further tuning. The resulting importance weights are clipped to $0 \le \widehat{w}(x) \le W$, with $W = 2$. No target outcome labels are used in fitting, tuning, or applying the density-ratio estimator.

This density-ratio estimator is part of the A-PAIR implementation. The CPCS baseline uses its own score-space density-ratio implementation described in App. G.2.

F.3 Frozen Objects, Source Screening, and Finite Classes of Audit Functions

All candidate values of θ_{anc} and all path iterates compared on the final holdouts are included in the simultaneous family. Initializers, path thresholds, binning rules, candidate paths, and audit functions are fixed using preprocessing and construction data before either independent audit holdout is evaluated. For every candidate predictor, repair-bin boundaries are likewise fixed before holdout evaluation. This permits reuse of a single source audit set and a single target audit set across the finite candidate family while accounting for the resulting comparisons through simultaneous audit bounds.

The finite-sample theory in Sec. 5 permits any valid simultaneous concentration bounds for the fixed finite classes of audit functions and any valid simultaneous one-sided bound for the paired source-Brier difference. Hoeffding or empirical-Bernstein bounds provide distribution-free finite-sample instantiations.

Audit instantiation used in the experiments. For the reported experiments, the total audit error budget is $\delta = 0.05$, divided equally among the anchor, preservation, repair, and source-Brier audits: $\delta_{\mathrm{stage}} = 0.0125$. For the anchor and preservation audits, each fixed direction j is assessed using the studentized Bonferroni quantity $U_j = |\widehat{\mu}_j| + t_{1-\delta_{\mathrm{stage}}/(2M),\, n-1}\widehat{\mathrm{SE}}_j$, where M is the number of directions in the corresponding finite class of audit functions. The repair audit uses a normal-approximation bound, and the source-Brier screen uses a paired normal approximation. These bounds are practical approximations used for the reported experiments and are distinct from the distribution-free finite-sample bounds used in the theoretical results.

Finite audit classes and approximate audit procedures. For infinite bounded audit classes, analogous results can instead be obtained using suitable pseudodimension or Rademacher-complexity bounds; this changes the concentration radius rather than the underlying audited population quantity. The implemented procedure uses finite classes of audit functions.

In particular, the preservation audit uses the 522 extreme points of a fixed degree-two anchor-score constraint polytope. The repair class is likewise finite and is formed from prespecified base functions crossed with current-score intervals. The exact class sizes used for each benchmark are reported in App. G.2. More general bounded Lipschitz or RKHS multiplier classes would require their own uniform-complexity control and are not used in the reported experiments.

Table 3: Per-run sample sizes. MMLU-Pro entries are ranges across shifted subjects.

Dataset	Source construction	Source audit	Target construction	Target audit	Target evaluation
MMLU-Pro	460–812	471–819	650–1,146	387–682	665–1,174
SciQ	1,413	1,413	1,507	1,507	1,507

G Experimental Details

A-PAIR path construction, auditing, and predictor selection were performed on the CPU of a Mac Studio equipped with an Apple M1 Ultra (20 CPU cores) and 64 GB of unified memory, running macOS 15.1.1.

G.1 Models, Distribution Shifts, and Data Splits

We use `Qwen2.5-1.5B` as the predecessor model and `Qwen2.5-1.5B-Instruct` as its fine-tuned counterpart on both benchmarks (Qwen et al., 2024). For each question, let $\hat{j}$ be the answer selected by the fine-tuned model. All methods predict the probability of the binary event $Y = \mathbf{1}\{\hat{j} \text{ is correct}\}$. The fine-tuned confidence q_{ft} and predecessor confidence q_{pre} are both evaluated on this same option $\hat{j}$. We compute model scores without free-form generation. For each valid answer option, we evaluate the conditional log probability of its answer-letter continuation and normalize the resulting scores over the valid options. The maximum sequence length is 2048.

For MMLU-Pro, we use 12,032 questions from 14 broad subjects. In each shift condition, one subject s has probability mass $P_{\mathrm{src}}(S = s) = 0.04$ and $P_{\mathrm{tgt}}(S = s) = 0.16$, with the remaining mass distributed uniformly across the other 13 subjects. We use three fixed splits for each shifted subject, giving 42 runs. Reported MMLU-Pro results are first averaged over the three runs within each subject and then equally over the 14 subjects.

For SciQ, we use 13,635 valid questions after preprocessing. We define the label-independent discrepancy $d(x) = |\mathrm{logit}\, q_{\mathrm{ft}}(x) - \mathrm{logit}\, q_{\mathrm{pre}}(x)|$ and divide the examples into five discrepancy quintiles. The highest-discrepancy quintile has mass 0.08 in the source and 0.32 in the target. The remaining four quintiles each have source mass 0.23 and target mass 0.17. We report averages over 30 deterministic hash splits.

For A-PAIR, labeled source data are divided into a construction set and an independent audit set. Cross-validation and path construction use only the construction set. The source audit set is used for anchor auditing, repair auditing, and the source-Brier screen. Target data are divided into construction, audit, and evaluation sets. Target construction and audit labels are not used. The target evaluation labels are accessed only after the anchor, stable/drift partition, candidate paths, and final predictor have been fixed. The per-run sample sizes are summarized in Tab. 3. MMLU-Pro sample sizes vary across shifted subjects because the 14 subject pools have different cardinalities; stratified sampling without replacement under the fixed source/target mixture proportions therefore yields different feasible per-run sample sizes. SciQ uses fixed sample sizes in every run.

G.2 Calibration Methods and Audit Protocol

In the reported experiments, we use $q_\theta^{(0)}(x) \equiv 0$ for every candidate θ. Density-ratio estimates are clipped at $W = 2$, and calibration-path computations are performed in double precision. We consider two predecessor-to-fine-tuned logit maps: the identity map and a fitted positive-slope affine map. The stable region of each candidate requires agreement between the predecessor and fine-tuned argmax answers and preferentially includes examples with smaller aligned logit discrepancy. For each anchor map, we construct candidate stable regions with stable-mass caps 0.05, 0.10, 0.20, 0.35, and 0.50. Candidates must pass the anchor audit. Among the passing candidates, we select the candidate $\hat{\theta}$ with the largest coverage on the unlabeled target audit set. If no candidate passes the anchor audit, A-PAIR uses the all-repair candidate.

The subgroup functions g begin with a 29-group score-based set: the constant function; quartile indicators for q_{pre}, q_{ft}, and the predecessor–fine-tuned discrepancy; and the cells of a 4×4 predecessor-

by-fine-tuned score grid. For MMLU-Pro, we additionally include 14 subject-category indicators, giving 43 subgroup functions. For SciQ, we instead include eight question-prefix indicators, giving 37 subgroup functions.

For repair, we augment these subgroup functions with q_{pre}, q_{ft}, and smooth directions derived from the score pair, giving 69 base functions for MMLU-Pro and 63 for SciQ. Each base function is crossed with 11 current-score intervals: the full interval $[0, 1]$ and ten equal-width bins $[0, 0.1), \ldots, [0.9, 1]$.

At each path-construction round, we compute the largest preservation and repair correlations and apply at most one update. When both directions exceed their thresholds, we select the update with the larger guaranteed decrease, $(c_{\mathrm{pres}})^2$ or $(c_{\mathrm{rep}})^2/W$. Predictions are clipped to $[0, 1]$ after each update. Each experimental path is capped at $T = 50$ updates and stops earlier if neither correlation exceeds its threshold. The preservation class contains 522 degree-two anchor-score directions.

We tune the repair stopping coefficient γ_0 over $\Gamma_0 = \{0.03, 0.02, 0.01, 0.005, 0.003, 0.001\}$ using three-fold cross-validation on the source construction set and the one-standard-error (1-SE) rule. Let $\gamma_\star$ be the selected coefficient and γ_{child} the next smaller value in Γ_0. The implementation uses γ_{child} to set the repair threshold γ_{rep}. We use anchor tolerance $\gamma_{\mathrm{anc}} = 0.05$ and preservation threshold $\gamma_{\mathrm{pres}} = 10^{-5}$.

The total experimental audit budget is $\delta = 0.05$, divided equally among the anchor, preservation, repair, and source-Brier audits, giving $\delta_{\mathrm{stage}} = 0.0125$. For the anchor and preservation audits, a fixed direction j is assessed using $U_j = |\widehat{\mu}_j| + t_{1-\delta_{\mathrm{stage}}/(2M),\, n-1}\widehat{\mathrm{SE}}_j$, where M is the size of the corresponding finite class of audit functions. Repair auditing and the paired source-Brier screen use normal approximations. For the selected candidate $\hat{\theta}$, $C_{\mathrm{obs}}(q) = \max\{U_{\mathrm{anc}} + U_{\mathrm{pres}}(q),\ U_{\mathrm{rep}}(q)\}$ is the final observable criterion. The selected predictor minimizes this criterion among candidates that pass the source-Brier screen. These studentized bounds are practical approximations used in the reported experiments, rather than the distribution-free bounds used in the theoretical analysis.

Raw-FT uses the fine-tuned confidence without post-hoc calibration. Calibrated Prediction with Covariate Shift (CPCS) transfers source-side calibration to a target distribution under covariate shift using importance weighting (Park et al., 2020). Our implementation estimates the target-over-source density ratio using L_2-regularized logistic regression on standardized model-score features. We use five-fold cross-fitting on the source data, regularization parameter $C = 1$, and the L-BFGS solver. This is a score-space implementation of the CPCS principle. It is not a byte-for-byte reproduction of a pipeline using the original paper's feature-space domain classifier.

For CPCS, we fit the temperature T_{temp} by minimizing the importance-weighted multiclass Brier loss over the answer-option distribution, using all labels from both the source construction and source audit sets. No target labels are used. We search 97 logarithmically spaced points over $T_{\mathrm{temp}} \in [10^{-3}, 10^{3}]$, followed by bounded scalar refinement. If the optimum lies at a boundary, we expand the range once to $[10^{-6}, 10^{6}]$ and repeat the search.

Disagreement-Aware Confidence Alignment (DACA) uses paired predictions from a predecessor model and its post-trained counterpart to perform confidence calibration without target correctness labels (Luo et al., 2025). Our implementation uses examples in the unlabeled target construction set for which the predecessor and fine-tuned argmax answers agree. We select the temperature that minimizes the mean KL divergence between the predecessor's full answer-option distribution and the temperature-scaled fine-tuned distribution on these examples. The search uses 160 logarithmically spaced temperatures from 0.05 to 20; if there are no agreement examples, we use $T = 1$. We include DACA because, like A-PAIR, it retains access to predictions from the predecessor model. Unlike A-PAIR, DACA does not use correctness labels.

G.3 Evaluation Metrics and Additional Diagnostics

We report target calibration error (TCE), Brier score, expected calibration error (ECE), smooth calibration error (smCE), and AUROC. TCE is the largest target calibration residual over the dataset-specific subgroup functions defined in App. G.2, crossed with the same 11 score intervals: the full interval $[0, 1]$ and ten equal-width bins. The additional score-valued and smooth directions used for repair are not included in the TCE evaluation class. The subgroup definitions and interval boundaries

Table 4: Target performance under label-restricted selection (mean ± nominal SE over 42 MMLU-Pro runs and 30 SciQ runs). Lower is better for TCE, Brier, ECE, and smCE; higher is better for AUROC. Bold marks the best mean within each dataset. Nominal SE is the sample SD across runs divided by $\sqrt{n}$.

MMLU-Pro Method	TCE	Brier	ECE	smCE	AUROC
Raw-FT	0.2334 ± 0.0021	0.241779 ± 0.001081	0.2341 ± 0.0021	0.2335 ± 0.0021	0.7190 ± 0.0031
CPCS	0.0271 ± 0.0024	0.181229 ± 0.001088	0.0378 ± 0.0022	0.0215 ± 0.0028	0.7190 ± 0.0029
DACA	0.0237 ± 0.0012	0.180268 ± 0.001041	$\mathbf{0.0371} \pm 0.0014$	0.0189 ± 0.0019	0.7193 ± 0.0029
A-PAIR	$\mathbf{0.0207} \pm 0.0012$	$\mathbf{0.178757} \pm 0.001100$	0.0379 ± 0.0019	$\mathbf{0.0159} \pm 0.0016$	$\mathbf{0.7228} \pm 0.0027$
SciQ Method	TCE	Brier	ECE	smCE	AUROC
Raw-FT	0.0083 ± 0.0003	0.013876 ± 0.000306	0.0094 ± 0.0003	0.0082 ± 0.0003	$\mathbf{0.9722} \pm 0.0022$
CPCS	$\mathbf{0.0042} \pm 0.0003$	0.013298 ± 0.000278	$\mathbf{0.0074} \pm 0.0005$	$\mathbf{0.0028} \pm 0.0003$	0.9718 ± 0.0023
DACA	0.0103 ± 0.0003	0.013331 ± 0.000265	0.0144 ± 0.0004	0.0101 ± 0.0003	0.9716 ± 0.0023
A-PAIR	0.0061 ± 0.0005	$\mathbf{0.013276} \pm 0.000265$	0.0084 ± 0.0005	0.0056 ± 0.0005	0.9603 ± 0.0039

are fixed before evaluation and shared across all methods; for each method, score-bin membership is determined by that method's predicted score. ECE uses 10 bins, and smCE is computed over $[-1, 1]$-valued 1-Lipschitz functions.

Tab. 4 supplements the mean comparison in Tab. 1 with nominal standard errors across runs.

We say that inheritance is active when the selected candidate $\hat{\theta}$ has positive stable coverage on the unlabeled target audit set, $\widehat{P}_{\mathrm{tgt,ho}}\{s(X) = 1\} > 0$. We call inheritance nontrivial when the stable coverage is at least 0.10. This is a property of the selected stable/drift partition and does not depend on the number of preservation updates. Inheritance is active in all 42 MMLU-Pro runs. Stable coverage is at least 0.10 in 40 runs, and its mean is 0.186. On SciQ, inheritance is active and nontrivial in all 30 runs, with mean stable coverage 0.504. Thus, the selected A-PAIR predictors consistently use predecessor-derived stable regions for inheritance.

The fine-tuned model has approximately 98% accuracy on the SciQ target evaluation sets, corresponding to 29.4 errors among 1,507 examples on average. In this low-error regime, scalar temperature calibration is already strong. An unweighted source-fitted temperature achieves TCE 0.004427, Brier score 0.013310, and smCE 0.002960, close to CPCS at 0.004222, 0.013298, and 0.002807, respectively. Their mean fitted temperatures are also similar, at 1.355 and 1.365. This supports the interpretation that a low-dimensional score correction captures much of the SciQ calibration structure, while A-PAIR attains the lowest mean Brier score.

We examine whether the predecessor score remains associated with correctness among examples with nearly the same fine-tuned probability. We divide fine-tuned probability into intervals of width 0.01, split each interval at the median predecessor score, and retain intervals with at least ten examples in each group. We then reweight the two groups to have the same distribution across these intervals. The higher-predecessor group is more accurate by 8.86 percentage points on MMLU-Pro and 0.79 points on SciQ. On SciQ, the corresponding weighted error rates are 0.352% and 1.145%, showing that the small difference in accuracy is larger relative to the remaining errors. The results are similar with interval width 0.02: 8.74 points on MMLU-Pro and 1.30 points on SciQ. For MMLU-Pro, repeating the comparison within each subject gives differences of 6.49 and 6.74 points at widths 0.01 and 0.02, respectively. Because fine-tuned probabilities are only approximately matched within each interval, these results are descriptive and do not estimate differences at exactly equal fine-tuned probability.