MODELS AS EVIDENCE: WHEN COLLABORATION MAKES BIAS CORRECTABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Models reused to assess a population can inherit selection bias from the sources of their training data. We study how to recover the population rate, the positive-label rate before any source selects labels, when training data and selection logs are unavailable but each model's task and sources are recorded. Whether the population rate can be consistently recovered depends on how the models share sources. One additional model, trained only on existing sources, can make the rate consistently recoverable even when neither it nor the existing collection suffices alone. When several sources may be biased, we give an exact condition on these source records: with m labels per source, the minimax mean squared error is $\Theta(m^{-1})$ when the condition holds and otherwise stays bounded away from zero, even when every model's fitted probability is observed exactly. A collection that fails this condition needs unbiased labels growing linearly with m to reach that rate from a fixed batch of models. The condition extends to conditional predictions of models that approximate known source mixtures. When a reference task shares the selection, a criterion checkable before acquisition tells apart two candidates that match in source count, overlaps, and unbiased outputs: only one enables recovery. On human feedback under controlled selection, a candidate with worse predictions yields lower Brier loss once combined with existing models through source records, without calibration labels.

1 INTRODUCTION

When one model is unreliable, a natural response is to combine several: evidence from several models can correct individual errors (Dietterich, 2000; Wu et al., 2019). Whether combining helps, however, depends on where their data came from. A trained model is not only a predictor but also a summary of the data it was trained on: selection, source mixing, and finite training determine what information it retains. Several trained models can even help reconstruct their training data (Wang & Kurz, 2022). Models that share sources therefore share evidence and can inherit the same selection bias even when parameters, predictions, or tasks differ.

Consider models that score whether responses are helpful, reused to estimate what fraction of responses users would find helpful. If positive feedback was preferentially retained when the training data were collected, a high predicted rate may reflect either response quality or selection. We call the positive-label rate in the population, before any source selects labels, the *population rate*; in our example, it is the true helpful rate. Beyond assessment, it is the prior needed to correct predicted probabilities under label-dependent selection (Saerens et al., 2002; Lipton et al., 2018). Fitting the selected feedback perfectly does not by itself determine this rate. We study its recovery when training data and selection logs are unavailable but each model's task and sources are recorded; we call these the *source records*. Our question is: *which additional model supplies the evidence needed for consistent recovery, even if it uses only sources already in the collection?*

Figure 1 shows how the answer can hinge on which sources the models share. Suppose the same sources also label whether responses are harmless, and each source selects labels in the same way for both tasks; we call harmlessness the *reference task* and helpfulness the *target task*. At most one source favors positive labels, and because selection is shared, the reference models constrain it. Two sources, the circle and the diamond, however, always appear together in the reference models, which therefore cannot tell them apart. Consider the two settings: circle biased or diamond biased.

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.

They can imply different helpful rates, and the existing target model cannot tell them apart either: it contains the circle but not the diamond, so a 50% helpful rate inflated by the circle and a 56% rate that this model sees without bias give the same expected output. Every existing model in O thus has the same expected output over training samples in both settings (Figure 1b). Candidate G uses both sources, so whichever is biased, its expected output follows the same increasing function of the helpful rate, and the two settings now give different outputs, 56% and 61%. Candidate H, like the existing target model, uses only one and leaves the settings tied, although the two candidates are matched on paper (Figure 1c). Nor does individual accuracy reveal which candidate helps: in our human-feedback experiments, a candidate with worse predictions enables better joint recovery without calibration labels. In Figure 1, neither G nor the existing collection can consistently recover the rate alone; together they can, which is the collaboration in our title. A model's value as evidence thus lies in how its sources relate to the collection, which source records reveal before acquisition.

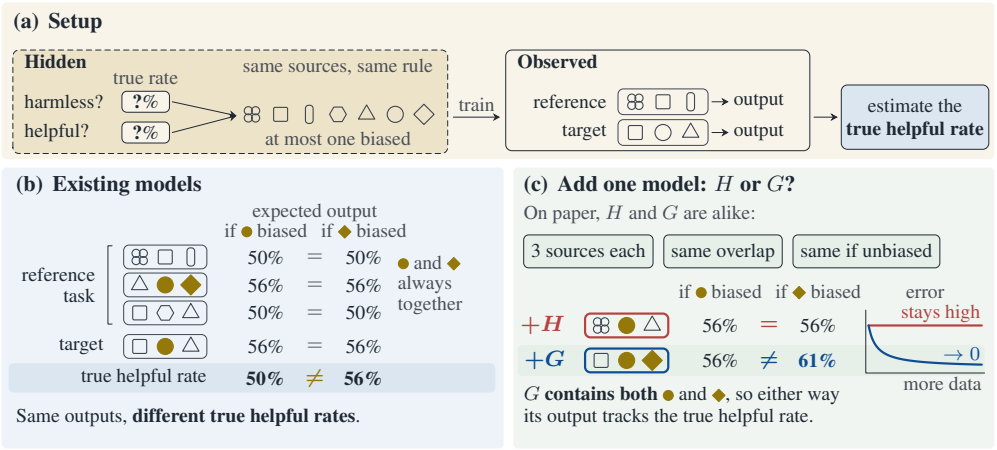

Figure 1: **One additional model, built from existing sources, can make a selection-biased rate recoverable.** (a) Shapes denote sources; boxes denote models trained on those sources. Only model tasks, sources, and outputs are observed. (b) Bias in the circle or diamond (settings A and B, Appendix C.2) gives every existing model the same expected output over training samples but different true helpful rates, $1/2$ and $5/9$. Percentages are rounded after undoing the known smoothing (Section 2). (c) H and G each use three existing sources and overlap equally with every existing model; without bias, the augmented collections have identical joint output distributions. With fitted probabilities observed, the curves sketch worst-case MSE as the per-source sample size m grows: $\Theta(1)$ for $O \cup H$ and $\Theta(m^{-1})$ for $O \cup G$ (Corollary 1).

The example shows how one model removes one ambiguity; consistent recovery requires removing every ambiguity that changes the target-task rate. This leaves two questions for the theory. Which relations among sources suffice for recovery? And when they are missing, can the variation of trained models across training samples offer another route? For the first question, we give exact conditions with matching risk bounds, including a criterion checkable before acquisition: it asks whether the added outputs can separate all population rates allowed by the bias family. For the second, we analyze the complete output distribution. The two settings of Figure 1 differ in this variation, but a more precisely trained model is still a single draw of it, so one batch of models cannot recover the rate consistently. Independent retraining can exploit distributional differences, but recovery is slow in the worst case near zero selection bias; adding G exposes target differences in the means. Replacing the missing model has a price: any collection that fails the recovery condition needs unbiased labels growing linearly with the per-source sample size to attain the parametric rate from a fixed batch. Our contributions are:

- **When recovery is possible.** We give exact conditions on the source records, with matching minimax risk bounds, for several biased sources in one task and for selection shared with a reference task, plus a sufficient condition for several shared signed biases. We also bound conditional-probability recovery for learned models approximating known source mixtures, and robustness to cross-task drift and incorrect records.

- **Which model to acquire.** With selection shared across tasks, a candidate enables recovery exactly when it resolves every ambiguous pair: two sources that the reference task cannot tell apart but that imply different target-task rates. This test uses only the source records. A matched construction shows that a collection and a candidate can each fail alone yet succeed together.
- **How resources compare.** A collection failing the recovery condition needs $\Omega(m)$ unbiased labels to reach the parametric rate, and in the two-task model we give matching risk bounds for retraining and single-label students.

2 Problem Setup

The difficulty in Figure 1 already arises within a single task: when several sources may be biased, their combined effects can mimic a change in the population rate. We therefore begin with one task and add the reference task in Section 3.3, where it supplies further constraints. Let $\theta \in I = [1/4, 3/4]$ be the population rate. Each of A sources, indexed by $a \in [A] = \{1, \dots, A\}$, independently samples from a population with this rate and applies its own label-dependent retention rule. If source a retains positive labels $\exp(b_a)$ times as often as negative labels, its retained rate is

$$q(\theta, b_a) = \frac{\theta \exp(b_a)}{1 - \theta + \theta \exp(b_a)}. \tag{1}$$

Selection thus adds b_a to the population log odds, as in case-control sampling (Prentice & Pyke, 1979). At most s sources are biased, with $|b_a| \le B$ for fixed $s \ge 1$ and $B > 0$; either direction of selection is allowed. The learner knows this family but not the biased sources or their strengths.

Each source supplies one block of m independent retained labels. Each of the L models, indexed by $i \in [L]$, pools the blocks of $J_i \ge 1$ sources; the source records form a matrix $C \in \{0,1\}^{L \times A}$ with row sums J_i. Whenever two models use the same source, they reuse its block, inheriting both its selection effect and its sampling variation. The configuration—the numbers of sources and models and which sources each model uses—remains fixed as sample sizes grow.

The downstream learner knows the source records, sample sizes, and training rule, but cannot access training records, source-level counts, or selection logs. The learner either observes each model's fitted positive-label probability directly or receives a student trained on n labels sampled from that probability, a hard-label form of distillation (Hinton et al., 2015; Lin et al., 2020). We call the former *complete observations* and write $n = \infty$. A student's parameter determines how many of its synthetic labels were positive (Appendix A.2), so each student reveals this count.

Each model fits a smoothed positive-label probability to its pooled labels. Undoing the known smoothing gives V_i, the empirical positive-label rate of model i's pooled labels, under complete observations; applied to a student's synthetic-label frequency, it gives an observation V_i that is conditionally unbiased for this rate (Appendices A and G.1). With $V = (V_i)_{i \in [L]}$, $W_{ia} = C_{ia}/J_i$, and $e = (e_a)_{a \in [A]}$, $e_a = q(\theta, b_a) - \theta$,

$$\mu := \mathbb{E}V = \theta \mathbf{1} + We, \qquad \|e\|_0 \le s, \qquad \mathrm{Var}(V_i) = O(m^{-1} + n^{-1}). \tag{2}$$

Here μ_i is the observation mean (the expected output in Figure 1), averaging over training samples and, for student observations, synthetic-label sampling. Which sources each model uses determines how the population rate and selection effects combine; m and n control observation noise. Define the minimax mean squared error for a collection with source records C,

$$\mathcal{R}_C(m, n) = \inf_{\widehat{\theta}} \sup_{\theta, b} \mathbb{E}_{\theta, b}(\widehat{\theta} - \theta)^2. \tag{3}$$

The supremum ranges over the stated family, and estimators may use all model observations. Consistency means vanishing worst-case risk as $m \to \infty$, and also $n \to \infty$ when observing students. This interface favors the learner, who knows the bias family, the weights W, and each model's fitted probability; our impossibility results therefore hold a fortiori whenever less is known.

Throughout, $\mathbf{1}$ and $\mathbf{0}$ are all-ones and all-zeros vectors, $\|v\|_0$ counts nonzero entries, and constants in $\lesssim$, $\asymp$, O, Ω, Θ may depend on the configuration and parameter bounds but not on sample sizes.

3 CONDITIONS FOR CONSISTENT RECOVERY

3.1 SEPARATING THE POPULATION RATE FROM SPARSE SOURCE BIAS

In $\mu = \theta\mathbf{1} + We$, changing θ while holding e fixed shifts every observation mean equally. Such a shift is invisible if a change in the source effects can reproduce it. Each parameter setting allows at most s biased sources, so comparing two settings can involve up to $2s$ sources. The separation margin measures how closely their effects can reproduce a unit shift in the population rate:

$$\kappa_s(W) = \min_{T \subseteq [A],\, |T| \le 2s} \operatorname{dist}_2(\mathbf{1}, \operatorname{col}(W_{:,T}))\,. \tag{4}$$

Here $W_{:,T}$ keeps the columns of W indexed by T, col is the column space (empty: $\{0\}$), and dist_2 is Euclidean distance to a subspace; $W_{\mathcal{C}}$ keeps the rows for a model collection $\mathcal{C}$. This adapts classical sparse-error separation (Fawzi et al., 2014) to recovery of θ. Although κ_s involves only observation means, Theorem 1 shows that it also decides recovery from the complete observation distribution.

Theorem 1 (Recovery with multiple biased sources)**.** *Under Section 2, consistent estimation is possible if and only if $\kappa_s(W) > 0$, with*

$$\mathcal{R}_C(m,n) \asymp \begin{cases} m^{-1} + n^{-1}, & \kappa_s(W) > 0, \\ 1, & \kappa_s(W) = 0, \quad m \ge m_0. \end{cases} \tag{5}$$

Both bounds allow any positive integer n and $n = \infty$, with $n^{-1} = 0$; the first holds for every $m \ge 1$. The threshold m_0 depends only on the fixed configuration and parameter bounds.

The condition determines the population rate without requiring the biased sources to be found. As in sparse-error separation for secure estimation (Fawzi et al., 2014), projecting the difference between two mean vectors away from the columns supporting their bias difference gives $|\theta - \theta'| \le \kappa_s(W)^{-1}\|\mu - \mu'\|_2$. Thus close observation means imply close population rates, even if the biased sources differ. Least-squares fitting over at most s source effects converts observation error into error in θ at the same order (Donoho, 1994), without knowing the selection strengths.

When $\kappa_s = 0$, we construct two settings with different population rates and identical observation means by splitting at most $2s$ sources into two admissible sets of biased sources. Trained models are random, so their complete count vector, whose covariance depends on the parameters, might still separate the settings. It does not: a uniform local limit approximation for overlapping lattice counts keeps the total variation distance between the two count laws below one as m grows (Appendix B.4). The obstruction is therefore statistical rather than an artifact of using means, giving a positive asymptotic risk lower bound (Tsybakov, 2009). Appendix G.1 proves both directions.

Acquisition appends a row to W. A candidate enables consistent recovery exactly when the augmented margin is positive, a condition checkable from source records alone. For every fixed s, a single added model can turn zero separation into positive separation while reusing existing sources and records; the candidate alone also fails to recover θ (Appendix G.2).

Additional models can also guard against errors in source records. With at most r models whose source records list incorrect sources, the parametric rate remains attainable if separation survives every deletion of up to $2r$ model rows. Such configurations exist for every fixed s, r (Appendix G.4).

3.2 CONDITIONAL PREDICTIONS FROM KNOWN SOURCE MIXTURES

Models are used for their predictions at individual inputs, not only for an average rate. The separation margin also works input by input, extending recovery to conditional probabilities when models approximate known mixtures of fitted source predictors. Let $p(x) = \Pr(Y = 1 \mid X = x)$ and let $q_a(x)$ be the conditional positive-label probability after source a's selection. At most s of the functions q_a differ from p, with input-dependent effects allowed. The population mixture is $\mu_i(x) = \sum_a W_{ia} q_a(x)$. The observed function $\widetilde{f}_i$ is either the fitted mixture $\sum_a W_{ia}\widehat{q}_a$ or a student distilled from it. With $e_a(x) = q_a(x) - p(x)$, we have $\mu(x) = p(x)\mathbf{1} + We(x)$. The learner uses the observed predictions and known weights W to estimate p.

Least-squares fitting at each input, with the fitted target clipped to $[0,1]$, gives, when $\kappa_s > 0$,

$$\mathbb{E}\|\widehat{p} - p\|^2_{L^2(P_X)} \le \frac{4}{\kappa_s(W)^2} \sum_i \mathbb{E}\|\widetilde{f}_i - \mu_i\|^2_{L^2(P_X)}, \tag{6}$$

where P_X is the target input distribution and the left side is expected excess Brier risk. The inverse margin controls an upper bound on how much the source structure amplifies approximation errors into error in p, so it ranks candidates before their outputs are seen; the approximation errors still depend on each candidate. These errors compare observed predictions with mixtures of the *post-selection* probabilities $q_a(x)$, not with p. A model that faithfully reproduces a biased mixture is therefore useful evidence, even though its own predictions are biased.

Their size depends on how accurately the source predictors are fitted, how well distillation preserves their mixtures, and how the distillation inputs cover P_X. Appendix H gives the bounds and a version selecting a common set of potentially biased sources across inputs; Section 5.2 tests it on human feedback.

3.3 Using a reference task and acquiring a matched candidate

When the target task alone cannot separate its rate from selection, a reference task that shares the selection supplies constraints that single-task models lack. Data-curation pipelines reuse filtering components, as documented in Dolma (Soldaini et al., 2024); motivated by this reuse, we study sources whose positive-to-negative retention ratios are shared across tasks. With a single, positive selection shared in this way, recovery is characterized by how pairs of sources appear in reference and target models, a condition that can be checked directly before acquisition. This result complements Theorem 1 rather than specializing it. Figure 1 is a concrete instance: reference-task models narrow the biased source down to the circle or the diamond, and the target-task models decide whether this ambiguity changes the target-task rate. Consider a reference task $t = 1$ and target task $t = 2$, with rates $\theta_t \in I$. Model i belongs to a known task t_i. At most one source is biased: $b_a = b\mathbf{1}\{a = a_\star\}$, with $b \in [0, B]$ for a fixed $B > 0$, shared across tasks ($a_\star$ is immaterial if $b = 0$). Every model uses $J \geq 2$ sources. Source–task blocks are independent, with reuse within each task as in Section 2, so

$$\mu_i = \theta_{t_i} + \frac{C_{ia_\star}}{J}[q(\theta_{t_i}, b) - \theta_{t_i}]. \tag{7}$$

We now write $\mathcal{R}_{\mathcal{C}}(m, n)$ for the minimax risk for θ_2 over this family. The collection contains a target model, and every source appears in some but not all reference models. Since selection only raises rates, reference models omitting the biased source attain the minimum $\min_{i:t_i=1} \mu_i = \theta_1$, and the range of reference means determines b. The reference task may still fail to reveal *which* source is biased, because sources with the same reference pattern (used by exactly the same reference models) are interchangeable there. Such a pair matters only when the target task covers one source fully and the other not at all:

$$\mathcal{U}(C) = \left\{(a, c) : C^{(1)}_{:,a} = C^{(1)}_{:,c}, \quad C^{(2)}_{:,a} = \mathbf{1}, \quad C^{(2)}_{:,c} = \mathbf{0}\right\}, \tag{8}$$

where $C^{(1)}, C^{(2)}$ are the reference and target rows of C. We call the pairs in $\mathcal{U}(C)$ *ambiguous*.

Theorem 2 (Recovery with shared cross-task selection)**.** *Under the setup of Section 3.3, $\mathcal{R}_{\mathcal{C}}(m, n) \asymp m^{-1} + n^{-1}$ if $\mathcal{U}(C) = \varnothing$. If $\mathcal{U}(C) \neq \varnothing$, then $\mathcal{R}_{\mathcal{C}}(m, n) \asymp 1$ for sufficiently large m, including complete observations.*

To see why this is the only obstruction, take two sources with the same reference pattern. A target model containing both or neither has the same strictly increasing mean function of θ_2 under either source choice (G in Figure 1c). If each source is absent from some target model, $\min_{i:t_i=2} \mu_i = \theta_2$ under either choice. Only full coverage versus absence permits different target-task rates. For different reference patterns, nearly equal observation means force both selection effects to be small. These facts bound differences in θ_2 uniformly by observation-mean differences (Appendix B).

A target-task candidate using a subset $\mathcal{S}$ of J existing sources resolves all ambiguous pairs in $\mathcal{U}(C)$ precisely when

$$\forall (a, c) \in \mathcal{U}(C), \qquad a \notin \mathcal{S} \quad \text{or} \quad c \in \mathcal{S}. \tag{9}$$

The candidate must omit a (in every existing target model) or include c (in none).

For the existing collection O in Figure 1, the circle and diamond form the only ambiguous pair: they share a reference pattern, and only the circle appears in the existing target model. G includes the diamond and satisfies (9); H does not.

Corollary 1 (Matched candidates and joint recovery)**.** *Within the parameter family of Section 3.3, the risk for $O \cup G$ observed through students trained on n synthetic labels each is $\Theta(m^{-1} + n^{-1})$. Complete model observations satisfy, as m grows,*

$$\mathcal{R}_O(m, \infty) \asymp \mathcal{R}_G(m, \infty) \asymp \mathcal{R}_{O \cup H}(m, \infty) \asymp 1, \qquad \mathcal{R}_{O \cup G}(m, \infty) \asymp m^{-1}. \tag{10}$$

Appendix C verifies the matching and individual lower bounds. On $O \cup G$, allowing the two tasks' selection log-odds shifts to differ by at most ε, with $0 \le \varepsilon \le \min\{B, 1\}/4$, gives matching risk $\Theta(m^{-1} + n^{-1} + \varepsilon^2)$ (Appendix G.3). Theorem 2 extends to known monotone selection channels and unequal source counts, reference models that separate several signed shared biases also suffice, and choosing several candidates at minimum cost is a weighted set cover (Appendix F).

4 Retraining, Unbiased Labels, and Students

With an ambiguous pair unresolved, increasing the per-source sample size cannot make recovery consistent (Theorem 2). What can replace the missing model? The two settings of Figure 1, circle or diamond biased, share observation means but not variances; each trained model is one draw of this variation, and retraining supplies more draws.

We therefore allow R independently retrained batches; drawing more synthetic labels from a fixed model would only approach complete observations (Appendix A.4). All batches share the same unknown parameters and configuration, but each uses fresh, independent samples from the sources; within a batch, models share sources as before. Beyond these R batches, the estimator may adaptively choose a task and draw an independent label from its population before selection. At most k such unbiased labels may be acquired in total. Write $\mathcal{R}_{\mathcal{C}}(m, n; R, k)$ for the resulting minimax risk.

Theorem 3 (Risk with retraining and unbiased labels)**.** *For the configuration in Figure 1 and the parameter family of Section 3.3, the minimax mean squared error bounds in Table 1 hold uniformly over all integers $m, R \ge 1$ and $k \ge 0$.*

Table 1: **Minimax risk with retraining and unbiased labels.** The budget k is the total number of additional unbiased labels across all R independent training batches.

Models per batch	Complete observations ($n = \infty$)	One synthetic label per model per batch ($n = 1$)
O alone	$\Theta((\sqrt{R} + k)^{-1})$	$\Theta((1 + k)^{-1})$
G alone	$\Theta((\sqrt{R} + k)^{-1})$	$\Theta((1 + k)^{-1})$
$O \cup G$	$\Theta((Rm + k)^{-1})$	$\Theta((R + k)^{-1})$

The $O \cup G$ row holds for every configuration without ambiguous pairs, and the O row for every configuration with one target model and an ambiguous pair; with several target models, retraining can be faster, so its rate is itself a property of the source records (Appendices D.3 and D.5). First set $k = 0$ to examine the different roles of m and R. For a target model using J sources, let S be its number of positive training labels, $\mu = \mathbb{E}S/(Jm)$ its expected empirical positive-label rate, and $\delta = \mu - \theta_2 \ge 0$ the increase due to selection. Its mean and variance satisfy

$$\frac{\mathrm{Var}(S)}{Jm} = \mu(1 - \mu) - (J - 1)\delta^2. \tag{11}$$

Thus, even when the mean cannot separate θ_2 from selection, independent retraining uses variation across training samples to estimate δ^2 from the count variance and recover θ_2. Estimating the scaled variance $\mathrm{Var}(S)/(Jm)$ from R batches has standard error of order $R^{-1/2}$.

The difficulty lies near zero bias, where variance changes only to second order (Ho & Nguyen, 2019; Heinrich & Kahn, 2018). Shrink the selection strength in the two Figure 1 settings until their target-task rates differ by h: for O, observation means coincide, and scaled count variances differ by order h^2. Equating this signal with the estimation error $R^{-1/2}$ gives $h \asymp R^{-1/4}$, and hence MSE $R^{-1/2}$.

After adding G, the same two settings differ in an observation mean by order h. Averaging across R batches gives standard deviation of order $(Rm)^{-1/2}$, so matching the signal to this noise yields

$h^2 \asymp (Rm)^{-1}$. Adding G turns a second-order variance signal into a first-order mean signal. Binomial approximation bounds (Zacharovas, 2019) give matching lower bounds (Appendix D).

Distillation compresses the observation further: a student trained on sampled labels is a randomized function of the fitted probabilities, so it cannot improve the optimal risk (Blackwell, 1953). A single synthetic label drawn from a random fitted probability P has marginal law $\mathrm{Ber}(\mathbb{E}P)$. In the two settings above, the reference fitted probabilities in O have the same joint distribution, the target-task model's fitted probability has the same expectation, and the two tasks use independent data. One synthetic label per model therefore makes the entire observation distribution identical, which retraining cannot undo. With G, the mean difference survives single-label sampling.

Two or more synthetic labels retain $\mathbb{E}[P^2]$, and hence the variance signal: with $R \geq 2$ batches, the target model has risk at most $O(\min\{1, (1+m/n)/\sqrt{R}\})$, matching the complete-observation rate $R^{-1/2}$ when $n \propto m$ (Appendix D.9).

The cost of unbiased labels follows the same structural boundary as recovery in Theorem 1.

Theorem 4 (The price of a missing model)**.** *Under Section 2, augment one fixed batch of complete model observations with k independent unbiased labels. The resulting minimax risk satisfies*

$$\mathcal{R}_C(m, \infty; 1, k) \asymp \begin{cases} (m+k)^{-1}, & \kappa_s(W) > 0, \\ (1+k)^{-1}, & \kappa_s(W) = 0, \quad m \geq m_0. \end{cases} \tag{12}$$

Both cases hold uniformly over $k \geq 0$; the first holds for all $m \geq 1$, and m_0 is independent of k.

The upper bounds combine the estimator of Theorem 1 with the label mean; the difficulty is a lower bound uniform in k. It lets equal-mean settings approach zero bias, where their count laws overlap uniformly (Appendix D.7). A missing model therefore has a label cost: a collection with $\kappa_s = 0$ needs $k = \Omega(m)$ unbiased labels to match the parametric risk attainable after adding a model that makes κ_s positive. The two-task analogue replaces $\kappa_s > 0$ with $\mathcal{U}(C) = \varnothing$ (Appendix D.8).

5 Experiments

We first examine how shared sources affect recovery and acquisition, then test on real human feedback whether a weaker predictor can be the more useful collaborator. Rate RMSE is in percentage points; Brier scores are scaled by 10^3. Appendix E gives protocols and additional results.

5.1 Model acquisition and label budgets

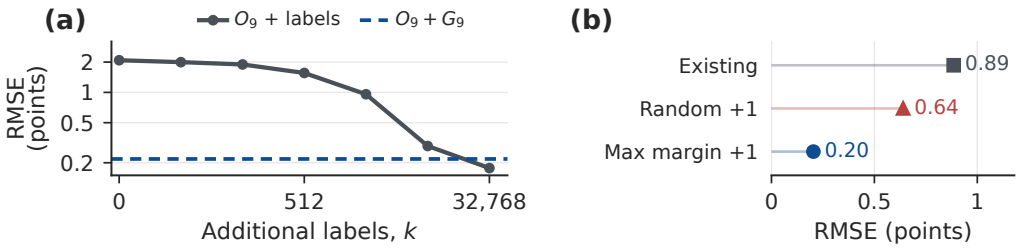

Figure 2: **Acquiring a model can substitute for collecting labels.** (a) At $m = 8192$, the solid curve adds k unbiased labels to O_9; the dashed line adds G_9 without labels. RMSE pools squared errors from two settings with 1,024 replicates each; model observations stay fixed across k within each replicate. (b) At $m = 16{,}384$, $n = 4m$, errors pool ten configurations and six parameter settings each; Table E.3 gives all sample sizes.

For the matched pair of Figure 1, increasing m from 32 to 8,192 reduces rate RMSE for $O \cup G$ from 5.46 to 0.40 points, while $O \cup H$ stays near 2.80, close to a Bayes reference that knows both settings (Table E.1). To illustrate Theorem 4, we then compare acquiring a model with collecting unbiased labels. Seven existing models O_9 each pool three of nine sources; candidate G_9 reuses these source

blocks and changes κ_2 from zero to positive (Table H.1). We use two fixed parameter settings with identical observation means for O_9 but different population rates. Generalized least squares fits the rate and two source effects without knowing which sources are biased.

At $m = 8192$, O_9 has rate RMSE 2.09 points, and adding G_9 without labels gives 0.22 (Figure 2a). On our grid, unbiased labels first match it at $k = 4m$ for every $m \geq 512$, whereas $k = m$ leaves 1.3–2.1 times its RMSE (Table E.2): the budget that replaces one model grows linearly in m.

We also test acquisition across 12 pre-specified single-task configurations with two biased sources. Ten initially have $\kappa_2 = 0$; each can reach $\kappa_2 > 0$ with one additional model built from existing sources. We select the candidate maximizing this margin from source records alone. Across new parameters and sampling replicates, at $m = 16{,}384$ and $n = 4m$, aggregate RMSE is 0.89 points for the existing collection, 0.64 for uniform random acquisition, and 0.20 for margin-based selection (Figure 2b). Both rules share the candidate set and estimator. A positive margin guarantees consistency over the bias family, whereas the RMSE averages over sampled parameters.

5.2 A WEAKER PREDICTOR CAN BE A BETTER COLLABORATOR

Moving to trained models, we implement the prediction interface of Section 3.2 with models trained on HelpSteer3 human feedback (Wang et al., 2025) and saved before acquisition. A label is positive when most annotators rate a response as mostly or perfectly helpful. Each source fits a logistic predictor to $m = 128$ retained records on frozen `Qwen3-Embedding-0.6B` features (Zhang et al., 2025). Known combinations of these predictors are distilled into neural networks on unlabeled texts. The learner receives only their predictions and the source records.

We reuse the nine-source configuration above and compare G_9 with candidate H_9, also using three sources and having the same overlap with every existing model. Neither candidate introduces a new source or new training data. Adding H_9 leaves $\kappa_2 = 0$, so source records select G_9 before its predictions are inspected. Appendix H.4 gives two target functions that G_9 separates.

Two sources undergo label-dependent or input-and-label-dependent selection, in either a positive or negative direction, giving four pre-specified conditions. G_9 includes both affected sources, whereas H_9 includes only one. After acquisition, the same correction rule fits $p(x)$ and two source effects to the acquired outputs at each input. It selects the pair of biased sources minimizing the weighted residual on 128 separate unlabeled inputs, without calibration labels.

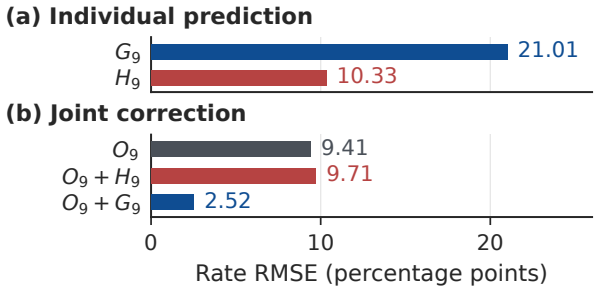

Figure 3: **A weaker predictor can be a better collaborator.** (a) Raw predictions from the two candidates. (b) The same source correction rule applied to each collection, without additional calibration labels. Errors average four selection conditions and 16 paired replicates on 1,024 new human-rated contexts.

On its own, G_9 has about twice H_9's rate RMSE: 21.01 versus 10.33 points. Yet the same correction rule gives 9.41 points with O_9, 9.71 with $O_9 \cup H_9$, and 2.52 with $O_9 \cup G_9$ (Figure 3). The three pre-specified Brier comparisons support the same reversal after familywise adjustment: corrected $O_9 \cup G_9$ outperforms corrected $O_9 \cup H_9$ and O_9, whereas raw G_9 performs worse than raw H_9 (Appendix E.4). The Brier gain comes mainly from correcting the average prediction over inputs.

Which sources the models share explains the reversal. If the affected sources were known, recovery would reduce to a simple cancellation. At a fixed input, let u and v denote the contributions of the two affected sources to a three-source mixture. The ideal predictions are $p + u$ and $p + v$ for two existing models, $p + u + v$ for G_9, and $p + u$ for H_9. Thus,

$$\underbrace{(p+u)+(p+v)}_{\text{two existing models}} - \underbrace{(p+u+v)}_{G_9} = p.$$

Same-sign effects move G_9's own prediction farther from p, yet supply the terms needed for cancellation. The learner does not know which sources are affected, hence which cancellation is valid;

the positive separation margin of $O_9 \cup G_9$ ensures that all sparse bias explanations consistent with its ideal predictions agree on p. Section 3.2 and Appendix H.2 bound how approximation and pair-selection errors degrade it.

Plain averaging of the $O_9 \cup G_9$ outputs gives 10.50 points, so the source records matter. Unbiased labels (Section 4) offer another route: Platt scaling (Platt, 2000; Guo et al., 2017) with 128 calibration labels brings $O_9 \cup H_9$ to 3.40 points. Appendices E.4–E.6 add controls and neural students.

6 Related Work

Model-reuse research describes individual models; we study how their sources relate. Learnware pairs trained models with specifications for reuse (Zhou, 2016), such as kernel mean embeddings of their training distributions (Wu et al., 2023), and density-assisted reuse corrects sample selection bias with input densities (Tang et al., 2023). Acquisition criteria value contributions to prediction or training: data valuation measures marginal contributions (Ghorbani & Zou, 2019) or replication-robust feature volume (Xu et al., 2021), and correlation-based client selection reduces training loss (Tang et al., 2022). We ask instead whether acquisition makes a population rate consistently recoverable.

Bias-correction methods differ in the extra information they use. Label-shift methods estimate target class proportions from calibrated source posteriors (Saerens et al., 2002) or from source label–prediction statistics under invariant class-conditional distributions (Lipton et al., 2018; Garg et al., 2020). Prediction-powered inference uses labeled observations of the target population (Angelopoulos et al., 2023). Dependency-aware ensembles, weak supervision, and confounder-aware evaluation use joint outputs on common items to learn dependencies and estimate reliability or latent quality (Jaffe et al., 2016; Bach et al., 2017; Zhao et al., 2026). We use source records together with model outputs, linked by known source mixtures and shared selection.

Sparse recovery supplies the structural tools: secure estimation gives the sparse-error separation and robust fitting we use (Fawzi et al., 2014), and optimal recovery and nonregular estimation relate error to inverse stability and distinguishability (Donoho, 1994; Ho & Nguyen, 2019; Heinrich & Kahn, 2018). Mixture decontamination recovers latent source distributions from direct samples of the mixtures (Katz-Samuels et al., 2019); we observe only models trained on them. Applied to our setting, these tools yield conditions on observation means; we show that mean confounding persists in the complete count distribution and compare how acquisition, retraining, and synthetic labels change the risk.

7 Discussion and Conclusion

Whether selection bias can be corrected depends on how models share sources, which individual accuracy need not reveal. We therefore argue that each model's tasks and sources should be recorded with it and used to screen candidates before acquisition; even a weaker predictor can help (Section 5.2). Labels can stand in for a missing model only at a cost: when the recovery condition fails, a fixed batch needs $\Omega(m)$ unbiased labels to reach minimax risk $O(m^{-1})$ (Theorem 4).

How a model is delivered also decides which corrections remain possible later: students trained on one synthetic label per model can erase the variance signal that retraining could exploit. The same question arises when models generate and filter their successors' training data, where retaining original data can change long-run outcomes (Shumailov et al., 2024; Gerstgrasser et al., 2024) and verifier-guided training can inherit the verifier's biases (Yi et al., 2026). Each stage thus preserves or destroys correction information; which stages can regain what an earlier one lost remains open.

Our guarantees assume a known bias family and known mixture weights; the latter also specifies what provenance records must contain. Across tasks, the characterization is exact for a single shared positive selection and sufficient for several signed ones, and a counterexample bounds where the slow retraining rate holds. The guarantees tolerate only a bounded number of incorrect records, yet audits (Longpre et al., 2024) and lineage reconstruction (Li et al., 2026) find such records often incomplete. Which records and model statistics jointly preserve recovery, including for models trained on mixed records, remains open. Extending these risk comparisons to sequential choices among acquisition, retraining, and direct measurement would let systems accumulate models while keeping the evidence needed to revise their conclusions.

AI USE STATEMENT

We used generative AI tools to improve the clarity and readability of the manuscript, assist with manuscript formatting and typesetting, assist with code writing, assist with drafting and checking parts of the mathematical proofs, and search for relevant literature and other research materials. All proofs were verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with AI assistance.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000. URL https://arxiv.org/abs/2301.09633.

Stephen H. Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the Structure of Generative Models without Labeled Data. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 273–282. PMLR, 2017. URL https://proceedings.mlr.press/v70/bach17a.html.

David Blackwell. Equivalent Comparisons of Experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

V. Chvátal. A Greedy Heuristic for the Set-Covering Problem. *Mathematics of Operations Research*, 4(3):233–235, 1979. doi: 10.1287/moor.4.3.233. URL https://pubsonline.informs.org/doi/abs/10.1287/moor.4.3.233.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by Exponential Weighting, Sharp PAC-Bayesian Bounds and Sparsity. *Machine Learning*, 72(1–2):39–61, 2008. doi: 10.1007/s10994-008-5051-0. URL https://arxiv.org/abs/0803.2839.

Thomas G. Dietterich. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pp. 1–15. Springer, 2000. doi: 10.1007/3-540-45014-9_1. URL https://doi.org/10.1007/3-540-45014-9_1.

David L. Donoho. Statistical Estimation and Optimal Recovery. *The Annals of Statistics*, 22(1):238–270, 1994. doi: 10.1214/aos/1176325367. URL https://web.stanford.edu/dept/statistics/cgi-bin/donoho/wp-content/uploads/2018/08/SEOR.pdf.

Hamza Fawzi, Paulo Tabuada, and Suhas Diggavi. Secure Estimation and Control for Cyber-Physical Systems Under Adversarial Attacks. *IEEE Transactions on Automatic Control*, 59(6):1454–1467, 2014. doi: 10.1109/tac.2014.2303233. URL https://arxiv.org/abs/1205.5073.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A Unified View of Label Shift Estimation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3290–3300, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/219e052492f4008818b8adb6366c7ed6-Abstract.html.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5B2K4LRgmz.

Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Philippe Heinrich and Jonas Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870, 2018. doi: 10.1214/17-aos1641. URL https://doi.org/10.1214/17-AOS1641.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, 2015. URL https://arxiv.org/abs/1503.02531.

Nhat Ho and XuanLong Nguyen. Singularity Structures and Impacts on Parameter Estimation in Finite Mixtures of Distributions. *SIAM Journal on Mathematics of Data Science*, 1(4):730–758, 2019. doi: 10.1137/18m122947x. URL https://arxiv.org/abs/1609.02655v4.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ariel Jaffe, Ethan Fetaya, Boaz Nadler, Tingting Jiang, and Yuval Kluger. Unsupervised Ensemble Learning with Dependent Classifiers. In *Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 351–360. PMLR, 2016. URL https://proceedings.mlr.press/v51/jaffe16.html.

Julian Katz-Samuels, Gilles Blanchard, and Clayton Scott. Decontamination of Mutual Contamination Models. *Journal of Machine Learning Research*, 20(41):1–57, 2019. URL https://jmlr.org/papers/v20/17-576.html.

Gregory F. Lawler and Vlada Limic. *Random Walk: A Modern Introduction*. Cambridge University Press, 2010. doi: 10.1017/cbo9780511750854. URL https://www.math.uchicago.edu/~lawler/srwbook.pdf.

Yu Li, Xiaoran Shang, Qizhi Pei, Yun Zhu, Xin Gao, Honglin Lin, Zhanping Zhong, Zhuoshi Pan, Zheng Liu, Xiaoyang Wang, Conghui He, Dahua Lin, Feng Zhao, and Lijun Wu. Tracing the Roots: A Multi-Agent Framework for Uncovering Data Lineage in Post-Training LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9606–9625. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.435. URL https://doi.org/10.18653/v1/2026.acl-long.435.

Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble Distillation for Robust Model Fusion in Federated Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2351–2363, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/18df51b97ccd68128e994804f3eccc87-Abstract.html.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and Correcting for Label Shift with Black Box Predictors. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and Yang Liu. Skywork-Reward-V2: Scaling Preference Data Curation via Human-AI Synergy. In *International Conference on Learning Representations*, pp. 133805–133838, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d8836d0294651a867c2ca51012923997-Abstract-Conference.html.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6 (8):975–987, 2024. doi: 10.1038/s42256-024-00878-8. URL https://www.nature.com/articles/s42256-024-00878-8.

John C. Platt. Probabilities for SV Machines. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large-Margin Classifiers*, pp. 61–74. MIT Press, 2000. doi: 10.7551/mitpress/1113.003.0008. URL https://doi.org/10.7551/mitpress/1113.003.0008.

R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979. doi: 10.1093/biomet/66.3.403. URL https://doi.org/10.1093/biomet/66.3.403.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446. URL https://doi.org/10.1162/089976602753284446.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. doi: 10.1038/s41586-024-07566-y. URL https://www.nature.com/articles/s41586-024-07566-y.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Anke Tang, Yong Luo, Han Hu, Fengxiang He, Kehua Su, Bo Du, Yixin Chen, and Dacheng Tao. Improving Heterogeneous Model Reuse by Density Estimation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4244–4252, 2023. doi: 10.24963/ijcai.2023/472. URL https://www.ijcai.org/proceedings/2023/472.

Minxue Tang, Xuefei Ning, Yitu Wang, Jingwei Sun, Yu Wang, Hai Li, and Yiran Chen. FedCor: Correlation-Based Active Client Selection Strategy for Heterogeneous Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10102–10111, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Tang_FedCor_Correlation-Based_Active_Client_Selection_Strategy_for_Heterogeneous_Federated_Learning_CVPR_2022_paper.html.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, 2009. doi: 10.1007/b13794. URL https://link.springer.com/book/10.1007/b13794.

Qian Wang and Daniel Kurz. Reconstructing Training Data From Diverse ML Models by Ensemble Inversion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2909–2917, 2022. URL https://openaccess.thecvf.com/content/WACV2022/html/Wang_Reconstructing_Training_Data_From_Diverse_ML_Models_by_Ensemble_Inversion_WACV_2022_paper.html.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh N. Sreedhar, and Oleksii Kuchaiev. HelpSteer 2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 1474–1501, 2024. doi: 10.52202/079017-0047. URL https://doi.org/10.52202/079017-0047.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Daniel Egert, Ellie Evans, Hoo-Chang Shin, Felipe Soares, Yi Dong, and Oleksii Kuchaiev. HelpSteer3: Human-Annotated Feedback and Edit Data to Empower Inference-Time Scaling in Open-Ended General-Domain Tasks. In *Proceedings of*

the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 25640–25662. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1246. URL https://doi.org/10.18653/v1/2025.acl-long.1246.

Xi-Zhu Wu, Song Liu, and Zhi-Hua Zhou. Heterogeneous Model Reuse via Optimizing Multiparty Multiclass Margin. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6840–6849. PMLR, 2019. URL https://proceedings.mlr.press/v97/wu19c.html.

Xi-Zhu Wu, Wenkai Xu, Song Liu, and Zhi-Hua Zhou. Model Reuse With Reduced Kernel Mean Embedding Specification. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):699–710, 2023. doi: 10.1109/tkde.2021.3086619. URL https://doi.org/10.1109/TKDE.2021.3086619.

Xinyi Xu, Zhaoxuan Wu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Validation Free and Replication Robust Volume-based Data Valuation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10837–10848, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/59a3adea76fadcb6dd9e54c96fc155d1-Abstract.html.

Bingji Yi, Qiyuan Liu, Yuwei Cheng, and Haifeng Xu. Escaping Model Collapse via Synthetic Data Verification: Near-term Improvements and Long-term Convergence. In *International Conference on Learning Representations*, pp. 24231–24260, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/29806b2e4dcb35e42abed21d43e2b7d8-Abstract-Conference.html.

Vytas Zacharovas. The Estimate of χ^2-Distance between Binomial and Generalized Binomial Distributions. *Theory of Probability and Its Applications*, 64(3):444–455, 2019. doi: 10.1137/S0040585X97T989593. URL https://arxiv.org/abs/1805.04891v2. Open manuscript: arXiv:1805.04891v2.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models, 2025. URL https://arxiv.org/abs/2506.05176.

Jitian Zhao, Changho Shin, Tzu-Heng Huang, Satya Sai Srinath Namburi GNVV, and Frederic Sala. CARE: Confounder-Aware Aggregation for Reliable LLM Evaluation, 2026. URL https://arxiv.org/abs/2603.00039v2. arXiv version 2, 8 September 2026.

Zhi-Hua Zhou. Learnware: on the future of machine learning. *Frontiers of Computer Science*, 10(4):589–590, 2016. doi: 10.1007/s11704-016-6906-3. URL https://journal.hep.com.cn/fcs/EN/10.1007/s11704-016-6906-3.

GUIDE TO THE APPENDIX

Appendix A specifies the training and observation rules. Appendix B proves the cross-task characterization in Theorem 2 and its acquisition condition (9); Appendix C verifies the matched construction and Corollary 1. Appendix D proves Theorems 3 and 4, extends the slow retraining rates to every configuration with one target model, and gives the cross-task label-budget bounds and the upper bound for student observations. Experimental methods and additional results appear in Appendix E, followed by extensions in Appendix F, including unequal source counts and several signed biases shared across tasks. Appendix G proves the sparse-bias classification in Theorem 1 and establishes robustness to cross-task drift and erroneous source records. Appendix H develops conditional-probability recovery and adaptive calibration.

Notation. Sources are indexed by $a \in [A]$ and models by $i \in [L]$. The source records form the matrix $C \in \{0,1\}^{L\times A}$. In the pooled-label model, the mixture matrix $W \in \mathbb{R}^{L\times A}$ has entries $W_{ia} = C_{ia}/J_i$; the conditional-prediction extension allows general known nonnegative mixture weights whose rows sum to one. Vectors indexed by models, such as V and μ, lie in $\mathbb{R}^L$; source-effect vectors $b = (b_a)$ and $e = (e_a)$ lie in $\mathbb{R}^A$. In the two-task model with a single biased source, b instead denotes the scalar selection strength. Subscripts on b_a and e_a index sources. In the two-task model, $t_i \in \{1,2\}$ is the task of model i, $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$, and parenthesized superscripts such as $C^{(t)}$ and $b^{(t)}$ index tasks. In the two-task model, $\eta = (\theta_1, \theta_2, a_\star, b)$ collects the parameters, and $\mu(\eta)$ is the mean vector in (7). For a collection $\mathcal{C}$ of models, $C_\mathcal{C}$ and $W_\mathcal{C}$ are the corresponding rows, and risks and margins indexed by $\mathcal{C}$ refer to these matrices: $\mathcal{R}_\mathcal{C} = \mathcal{R}_{C_\mathcal{C}}$ and $\kappa_s(W_\mathcal{C})$; the risk symbol $\mathcal{R}$ always refers to the parameter family of the section in which it appears. $M_{:,T}$ keeps the columns of a matrix M indexed by T, $M_{:,a}$ is column a, $\mathbf{1}\{\cdot\}$ is an indicator, and $\|f\|^2_{L^2(P)} = \mathbb{E}_{X\sim P}\|f(X)\|_2^2$. The scalar B bounds selection strength; m is the number of retained labels per source block, n the number of synthetic labels per student, R the number of independently retrained batches, and k the total number of additional unbiased labels.

The asymptotic notation defined at the end of Section 2 is used throughout; when a bound involves a local parameter such as h or ε, its constants are also uniform over the stated range of that parameter. $\mathrm{Ber}(p)$ and $\mathrm{Bin}(N,p)$ are the Bernoulli and binomial laws, $\overset{\mathrm{ind}}{\sim}$ indicates independent draws, $\mathcal{L}(X)$ is the law of X, and $\bigotimes$ forms product laws. For laws P, Q, $\mathrm{TV}(P,Q) = \sup_E |P(E) - Q(E)|$, $\mathrm{KL}(P\|Q) = \mathbb{E}_P \log(dP/dQ)$, $\chi^2(P\|Q) = \mathbb{E}_Q(dP/dQ - 1)^2$, and $H^2(P,Q) = \int(\sqrt{dP} - \sqrt{dQ})^2$; $\mathrm{kl}(p,q) = \mathrm{KL}(\mathrm{Ber}(p)\|\,\mathrm{Ber}(q))$. We write $\sigma(u) = (1+e^{-u})^{-1}$, $\mathrm{logit}(p) = \log[p/(1-p)]$, $\mathrm{clip}_{[a,b]}(x) = \min\{\max\{x,a\},b\}$, $[x]_+ = \max\{x,0\}$, $\|M\|_{\mathrm{op}}$ for the spectral norm, and $M^\dagger$ for the Moore–Penrose inverse.

A OBSERVATION MODEL AND TRAINING RULES

We detail the training and observation rules for the two-task design of Section 3.3. Write A for the number of sources, $t_i \in \{1,2\}$ for the task of model i, L_t for the number of task-t models, $L = L_1 + L_2$, and $I = [1/4, 3/4]$. The single-task experiment uses the same rules with J_i sources in model i, as specified in Appendix G.1. All constants below may depend on the fixed source records C, J, and B, but not on the sample budgets. Set $n^{-1} = 0$ for complete observations.

A.1 SELECTED SAMPLES AND SHARED DATA BLOCKS

Let $s_{a,y} > 0$ be the retention probability for label y at source a, with $s_{a,1}/s_{a,0} = \exp(b_a)$. Bayes' rule gives

$$\Pr(Y = 1 \mid \text{retained}, a, t) = \frac{s_{a,1}\theta_t}{s_{a,1}\theta_t + s_{a,0}(1-\theta_t)} = q(\theta_t, b_a). \tag{A.1}$$

Independent proposals and retention decisions imply that the first m retained labels are independent with this distribution. Denote them by $Y_{a\ell t}$, $\ell = 1, \dots, m$. Blocks for distinct source–task pairs are independent. The count for model i is

$$S_i = \sum_a C_{ia} \sum_{\ell=1}^{m} Y_{a\ell t_i}, \qquad N = Jm, \qquad T_i = S_i/N. \tag{A.2}$$

The same $Y_{a\ell t}$ occurs in every task-t model using source a.

A.2 Teacher and student training

The teacher fits a binary probability with one pseudocount for each label:

$$\mathcal{L}_i(p) = -(S_i + 1)\log p - (N - S_i + 1)\log(1 - p). \tag{A.3}$$

Strict convexity and the first-order condition give

$$P_i = \frac{S_i + 1}{N + 2}, \qquad w_i = \log\frac{P_i}{1 - P_i}, \qquad S_i = (N + 2)\sigma(w_i) - 1, \tag{A.4}$$

where $\sigma(u) = (1+e^{-u})^{-1}$. Thus the fitted probability, logit parameter, pooled count, and empirical rate determine one another.

To train a student, generate n labels independently conditional on P_i, with success probability P_i, and write $U_i = n^{-1}\sum_{j=1}^{n}\widetilde{Y}_{ij}$. Synthetic draws are conditionally independent across models. The student minimizes

$$F_i(v) = \log(1 + e^v) - U_i v + \frac{\lambda}{2}v^2, \qquad \lambda > 0. \tag{A.5}$$

Its unique minimizer satisfies $U_i = \sigma(v_i) + \lambda v_i$. The derivative of the right-hand side is strictly positive, so the student parameter and synthetic positive count are equivalent observations. Conditional on that count, the ordering of the synthetic labels is uniform and independent of the unknown parameters.

Define

$$\alpha_m = 1 + \frac{2}{Jm}, \qquad V_i = \alpha_m U_i - \frac{1}{Jm}. \tag{A.6}$$

Then $\mathbb{E}[V_i \mid S_i] = T_i$. Under complete observations we instead set $V_i = T_i$.

A.3 Means, variances, and dependence

Writing $d(p, b) = q(p, b) - p$, the transformed observations V_i satisfy

$$\mathbb{E}_\eta V_i = \theta_{t_i} + \frac{C_{ia_\star}}{J}d(\theta_{t_i}, b) = \mu_i(\eta). \tag{A.7}$$

Conditional variance decomposition yields

$$\operatorname{Var}_\eta V_i = \operatorname{Var}_\eta T_i + \frac{\alpha_m^2}{n}\mathbb{E}_\eta[P_i(1 - P_i)] \le \frac{1}{4Jm} + \frac{\alpha_m^2}{4n}. \tag{A.8}$$

The first term is bounded by summing the variances of the Jm independent labels in one model. For distinct models on the same task,

$$\operatorname{Cov}_\eta(T_i, T_j) = \frac{1}{J^2 m}\sum_a C_{ia}C_{ja}q(\theta_t, b_a)[1 - q(\theta_t, b_a)]. \tag{A.9}$$

For different tasks the covariance is zero. Conditional independence of synthetic draws and their conditional means imply $\operatorname{Cov}(V_i, V_j) = \operatorname{Cov}(T_i, T_j)$ for $i \neq j$.

A.4 Queries to fixed models

Write $Z_{\text{obs}} = (C, S)$ for the complete observations. Suppose a query history $\mathcal{H}$ is generated from Z_{obs} by a possibly adaptive random kernel that does not otherwise depend on η. Any estimator using $(Z_{\text{obs}}, \mathcal{H})$ can be simulated from Z_{obs} and independent randomization: generate the queries and responses sequentially using that same kernel. Conversely, an estimator can ignore $\mathcal{H}$. The two experiments therefore have the same optimal risks, as in classical comparisons of statistical experiments (Blackwell, 1953). In particular, a lower bound under complete model observations also applies to observations of students and to any such querying procedure. Additional synthetic labels can still improve estimates based on a student by reducing its sampling error.

B Proof of the Cross-Task Structural Characterization

This appendix proves Theorem 2 under the two-task assumptions of Section 3.3.

B.1 Global stability

For the logistic selection channel on $I \times [0, B]$, with any fixed $B > 0$, there are constants $c_q, C_q, D > 0$ such that

$$c_q \le \partial_b q(p, b) \le C_q, \qquad \partial_p q(p, b) > 0, \qquad |\partial_p d(p, b)| \le D, \qquad d(p, 0) = 0. \tag{B.1}$$

Indeed, $\partial_b q = q(1 - q)$ and, with $x = \exp(b)$, $\partial_p q = x/[1 + (x - 1)p]^2$, which is positive and bounded on this compact domain. The retained probabilities are uniformly bounded away from zero and one, which gives c_q. For example, $c_q = 3/(16e)$, $C_q = 1/4$, and $D = 1$ suffice for $B \le 1$; for larger B the constants depend on B.

Lemma B.1 (Global stability of the population rates)**.** *Write $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$ for the two task rates. If $\mathcal{U}(C) = \varnothing$, there exists $K < \infty$ such that, for every pair of admissible parameters,*

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty \le K \|\mu(\eta) - \mu(\eta')\|_\infty. \tag{B.2}$$

Proof. Let $e_0 = \|\mu(\eta) - \mu(\eta')\|_\infty$, and let a, c be the two candidate biased sources. Reference coverage and nonnegative selection imply

$$\theta_1 = \min_{i:t_i=1} \mu_i, \qquad d(\theta_1, b)/J = \max_{i:t_i=1} \mu_i - \min_{i:t_i=1} \mu_i. \tag{B.3}$$

Consequently $|\theta_1 - \theta_1'| \le e_0$, and the two reference ranges differ by at most $2e_0$.

If the reference columns for a, c differ, interchange the two parameters if needed to find a reference row with $(C_{ia}, C_{ic}) = (1, 0)$. Comparing that row with the two reference minima gives $d(\theta_1, b)/J \le 2e_0$. The other range is at most $4e_0$. Since $c_q b \le d(p, b) \le C_q b$, $b + b' \le 6Je_0/c_q$. Any target row then gives

$$|\theta_2 - \theta_2'| \le e_0 + (C_q/J)(b + b') \le (1 + 6C_q/c_q)e_0. \tag{B.4}$$

If the reference columns coincide, the range comparison gives

$$|b - b'| \le A_J e_0, \qquad A_J = (2J + D)/c_q. \tag{B.5}$$

Set $F_j(p, b) = p + (j/J)d(p, b) = (1 - j/J)p + (j/J)q(p, b)$ for $j \in \{0, 1\}$ and $a_0 = 1 - 1/J > 0$. Since $\partial_p q > 0$, these functions satisfy $\partial_p F_j \ge a_0$ and $|\partial_b F_j| \le C_q/J$. If the two target columns agree at some row, their common entry j therefore implies

$$a_0|\theta_2 - \theta_2'| \le e_0 + (C_q/J)|b - b'| \le [1 + (C_q/J)A_J]e_0. \tag{B.6}$$

Otherwise the target columns are complementary at every row. If both are nonconstant, each has a zero entry; their minimum target-task observation means are θ_2, θ_2', giving $|\theta_2 - \theta_2'| \le e_0$. The remaining case is a full column and a zero column, which would put one of $(a, c), (c, a)$ in $\mathcal{U}(C)$. It is excluded by hypothesis. Taking the maximum of these bounds proves the result, including at $b = 0$. □

B.2 An estimator attaining the upper bound

Choose a measurable minimum-residual fit over the compact parameter family:

$$\widehat{\eta} \in \arg\min_{\eta} \|V - \mu(\eta)\|_\infty. \tag{B.7}$$

The true parameter is feasible, so $\|\mu(\widehat{\eta}) - \mu(\eta)\|_\infty \le 2\|V - \mu(\eta)\|_\infty$. Lemma B.1, the inequality $\max_i z_i^2 \le \sum_i z_i^2$, and (A.8) give

$$\mathbb{E}_\eta \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_\infty^2 \le LK^2 \left(\frac{1}{Jm} + \frac{\alpha_m^2}{n} \right). \tag{B.8}$$

No independence across model observations is required.

An estimator that considers each possible biased source in turn also attains this order. For each $a \in [A]$, average the reference observations omitting a and project the result onto I, obtaining $\widehat{\theta}_{1,a}$. Let $z_a = J(\overline{V}_{\mathrm{ref},a} - \widehat{\theta}_{1,a})$, where the bar averages reference models containing a. Invert $d(\widehat{\theta}_{1,a}, b) = z_a$ on $[0, B]$, using the nearest endpoint when z_a lies outside its range. For a fixed target row i_0, invert $F_{C_{i_0 a}}(p, \widehat{b}_a) = V_{i_0}$ on I in the same manner. Select the resulting parameter configuration with smallest maximum residual across all model observations.

Both reference groups are nonempty. If each numerical inverse has parameter error at most τ, the true-source branch satisfies

$$|\widehat{\theta}_1 - \theta_1| \le e_V, \quad |\widehat{b} - b| \le A_J e_V + \tau, \quad |\widehat{\theta}_2 - \theta_2| \le K_1 e_V + Q\tau, \tag{B.9}$$

where $e_V = \|V - \mu(\eta)\|_\infty$, $K_1 = [1 + (C_q/J)A_J]/a_0$, and $Q = 1 + C_q/(Ja_0)$. The first bound follows from averaging and projection. For the second, the inverse residual is at most $(2J + D)e_V$ and its derivative is at least c_q. The third follows from the derivatives of F_j. Lipschitz continuity of μ then bounds the true branch's residual by $C(e_V + \tau)$. The selected branch has no larger residual, and Lemma B.1 gives

$$\mathbb{E}_\eta(\widehat{\theta}_{2,\mathrm{enum}} - \theta_2)^2 \le C'(m^{-1} + n^{-1} + \tau^2). \tag{B.10}$$

Bisection with $\tau = (m + n)^{-1}$, or $\tau = m^{-1}$ under complete observations, preserves the risk order. Appendix E.1 defines the GLS estimator used in the experiments.

B.3 Lower bounds from the two sample budgets

For distributions P, Q with target values separated by Δ, a test that selects the target closer to an estimate gives the standard two-point bound (Tsybakov, 2009)

$$\inf_{\widehat{\theta}} \max\{\mathbb{E}_P(\widehat{\theta} - \theta_P)^2, \mathbb{E}_Q(\widehat{\theta} - \theta_Q)^2\} \ge \frac{\Delta^2}{8}[1 - \mathrm{TV}(P, Q)]. \tag{B.11}$$

Indeed, a testing error costs at least $\Delta^2/4$, and the sum of testing errors is at least $1 - \mathrm{TV}(P, Q)$. We also use

$$\mathrm{KL}(\mathrm{Ber}(p) \| \mathrm{Ber}(q)) \le \frac{(p - q)^2}{q(1 - q)}. \tag{B.12}$$

For N_0 independent Bernoulli observations, take probabilities $1/2 \pm h$ with $h = (16\sqrt{N_0})^{-1}$. The product KL is at most $1/8$, hence total variation is at most $1/4$ by Pinsker's inequality. Equation (B.11) gives risk at least a constant times N_0^{-1}.

Apply this argument to the unbiased submodel, fixing θ_1. Even revealing all Am target-task source labels provides only Am independent Bernoulli observations about θ_2. Since model observations are generated from these labels, this proves the m^{-1} lower bound.

For the n^{-1} lower bound, one synthetic draw from (A.4) can be generated by sampling uniformly from the N training-label positions and two additional positions carrying the fixed labels $0, 1$. Reveal the sampled positions as well as the returned labels. Their distribution is parameter independent, and the $L_2 n$ target draws expose at most $L_2 n$ distinct source labels. A label reused by several models or draws is counted once. Conditional on the positions, the revealed distinct labels are independent Bernoulli observations in the unbiased submodel. The same KL bound therefore proves an n^{-1} lower bound. The larger of the two bounds is at least a constant times $m^{-1} + n^{-1}$, completing the first case of Theorem 2.

B.4 A uniform approximation on the observation lattice

The following lemma adapts lattice local limit arguments (Lawler & Limic, 2010, Chapter 2) to overlapping model counts, including linearly dependent model rows. It is stated on the lattice generated by the observations.

Lemma B.2 (Uniform lattice approximation). *Fix integer vectors $u_1, \dots, u_M \in \mathbb{Z}^L$. Let $Z_p = \sum_j u_j Y_j$ with independent $Y_j \sim \mathrm{Ber}(p_j)$ and $p_j \in [\epsilon, 1 - \epsilon]$, where $\epsilon > 0$ is fixed. Choose a basis of the lattice $\Lambda = \mathrm{span}_{\mathbb{Z}}\{u_j\}$ of integer combinations of the u_j, of rank $d \ge 1$, and write*

$u_j = H_\Lambda w_j$ *in its integer coordinates. Set*

$$W_p = \sum_j w_j Y_j, \quad \nu_p = \mathbb{E}W_p, \quad \Sigma_p = \sum_j p_j(1-p_j) w_j w_j^\top.$$

If $P_{p,m}$ *is the mass function of a sum of* m *independent copies of* W_p*, then*

$$\sup_p \sup_{z \in \mathbb{Z}^d} \left| m^{d/2} P_{p,m}(z) - \phi_{\Sigma_p}\left(\frac{z - m\nu_p}{\sqrt{m}}\right)\right| \longrightarrow 0, \tag{B.13}$$

$$\sup_p \sum_{z \in \mathbb{Z}^d} \left| P_{p,m}(z) - m^{-d/2} \phi_{\Sigma_p}\left(\frac{z - m\nu_p}{\sqrt{m}}\right)\right| \longrightarrow 0. \tag{B.14}$$

Here ϕ_Σ *denotes the centered Gaussian density with covariance* Σ*. For parameter pairs with* $\nu_p = \nu_{p'}$*, uniformly over such pairs,*

$$\mathrm{TV}(P_{p,m}, P_{p',m}) - \mathrm{TV}\{\mathcal{N}(0,\Sigma_p), \mathcal{N}(0,\Sigma_{p'})\} \longrightarrow 0. \tag{B.15}$$

Proof. Changing to lattice coordinates is a bijection on the observable support. The columns w_j generate $\mathbb{Z}^d$ and span $\mathbb{R}^d$. Thus, uniformly in p, $0 < \lambda_- I \preceq \Sigma_p \preceq \lambda_+ I$ in the positive-semidefinite order, for fixed constants λ_-, λ_+. The support contains 0 and each w_j, each with probability at least ϵ^M. If the characteristic function $\varphi_p(v) = \mathbb{E}e^{iv^\top W_p}$ has modulus one, these support points must all have equal phase. Hence $v^\top w_j \in 2\pi\mathbb{Z}$ for every j, which implies $v \in 2\pi\mathbb{Z}^d$. Compactness gives a uniform bound $|\varphi_p(v)| \le \rho < 1$ outside any fixed neighborhood of zero in $[-\pi, \pi]^d$.

The centered characteristic function $\psi_p(v) = e^{-iv^\top \nu_p}\varphi_p(v)$ satisfies, in a uniform neighborhood of zero,

$$\log \psi_p(v) = -\tfrac{1}{2} v^\top \Sigma_p v + O(\|v\|^3), \qquad |\psi_p(v)| \le e^{-c\|v\|^2}. \tag{B.16}$$

All remainder bounds are uniform because the support is fixed and the covariance eigenvalues have a positive lower bound. Fourier inversion after $v \mapsto v/\sqrt{m}$ gives

$$m^{d/2} P_{p,m}(z) = \frac{1}{(2\pi)^d} \int_{[-\pi\sqrt{m}, \pi\sqrt{m}]^d} e^{-iv^\top (z - m\nu_p)/\sqrt{m}} \psi_p(v/\sqrt{m})^m \, dv. \tag{B.17}$$

The Gaussian inversion formula has the same phase and factor $e^{-v^\top \Sigma_p v/2}$. On $\|v\| \le K_0$, the integrands converge uniformly. On $K_0 < \|v\| \le \delta\sqrt{m}$, a uniform Gaussian envelope bounds the integral. On the remaining region, the original characteristic function gives a bound of order $m^{d/2}\rho^m$. Letting m and then K_0 tend to infinity proves (B.13), uniformly in z since the phase has modulus one.

Within distance $K_0\sqrt{m}$ of the mean there are at most $c_{K_0} m^{d/2}$ lattice points, with c_{K_0} depending only on K_0 and d, so (B.13) implies convergence of the summed approximation error there. Outside this region, Chebyshev's inequality bounds the true mass by $\mathrm{tr}(\Sigma_p)/K_0^2$. Gaussian lattice tails tend uniformly to zero as K_0 grows: the covariance eigenvalues lie in a compact positive interval, and the fractional grid shift lies in $[0,1]^d$. This proves (B.14). For equal means, use the same grid and center for the two approximations. Their absolute difference is a Riemann sum converging uniformly to the integral of the absolute difference of the two Gaussian densities, proving (B.15). □

B.5 Nonvanishing risk when a source pair remains ambiguous

Fix $(a, c) \in \mathcal{U}(C)$ and choose a sufficiently small $b_* \in (0, B]$. Let $\Delta = d(1/2, b_*)/J > 0$ and consider

$$\eta_A = (1/2, 1/2, a, b_*), \qquad \eta_B = (1/2, 1/2 + \Delta, c, b_*). \tag{B.18}$$

Both parameters are admissible for small enough b_*. Interchanging the reference data blocks for a, c preserves all reference counts, so their joint laws are equal. In the target task, a appears everywhere and c nowhere; every model therefore has observation mean $1/2 + \Delta$ in both configurations.

The full count vector is a sum of m independent vectors with fixed integer loadings given by the task-specific membership columns. These loadings generate the same lattice under both parameters. All Bernoulli probabilities are uniformly interior. Lemma B.2 implies that the total variation distance

converges to a value $\tau_* < 1$, since both limiting Gaussian distributions have positive densities on the same space. Consequently, for all sufficiently large m, (B.11) gives

$$\mathcal{R}_{\mathcal{C}}(m, \infty) \geq \frac{\Delta^2(1-\tau_*)}{16} > 0. \tag{B.19}$$

Finite student observations inherit this lower bound by postprocessing. The constant estimate $1/2$ has risk at most $1/16$, proving the second case of Theorem 2.

B.6 Proof of the acquisition condition

Adding a target row does not change reference membership patterns. A pair $(a, c) \in \mathcal{U}(C)$ remains ambiguous exactly when the new row contains a and omits c. No new pair can enter $\mathcal{U}(C)$: adding a target row cannot turn a nonconstant column into a full or zero column. Hence all ambiguous pairs are resolved exactly under (9). The condition requires the candidate's task and source records; estimation after acquisition additionally uses its numerical observation.

C Matched Candidates and Joint Recovery

C.1 Source records and matching

Table C.1 specifies the configuration in Figure 1. Sources $1, \ldots, 7$ correspond respectively to the clover, square, circle, vertical capsule, diamond, hexagon, and triangle. The existing collection is $O = \{R_1, R_2, R_3, T_0\}$.

Table C.1: **Memberships underlying Figure 1.** The circle and diamond are sources 3 and 5.

Model	Task	Sources
R_1	Reference	$\{1, 2, 4\}$
R_2	Reference	$\{3, 5, 7\}$
R_3	Reference	$\{2, 6, 7\}$
T_0	Target	$\{2, 3, 7\}$
G	Target	$\{2, 3, 5\}$
H	Target	$\{1, 3, 7\}$

The reference equivalence classes are $\{1, 4\}, \{2\}, \{3, 5\}, \{6\}, \{7\}$. Only $(3, 5)$ belongs to $\mathcal{U}(C_O)$: source 3 occurs in the existing target model and source 5 does not. Candidate G resolves this ambiguity, whereas H leaves it unchanged. Both candidates use three sources and have overlap vector $(1, 2, 1, 2)$ with (R_1, R_2, R_3, T_0). All seven sources occur in O. Each candidate adds one previously unobserved source–target block: source 5 for G, source 1 for H.

When $b = 0$, the target count pair after either acquisition has the representation

$$(S_{T_0}, S_{\text{new}}) = (X_c + X_o, X_c + X_n), \tag{C.1}$$

where $X_c \sim \text{Bin}(2m, \theta_2)$ and $X_o, X_n \sim \text{Bin}(m, \theta_2)$ are independent. The reference count law is unchanged and independent of the target block. Thus the entire joint count distributions agree after identifying the new-model coordinate. Teacher training and the identical student sampling kernels preserve this equality for every m, n.

C.2 An explicit pair of parameter configurations

For $B \geq \log 2$, set $\theta_1 = 1/2$ and $b = \log 2$. The two settings used in the experiments are

$$\eta_A = (1/2, 1/2, 3, \log 2), \qquad \eta_B = (1/2, 5/9, 5, \log 2). \tag{C.2}$$

Their target-task observation means are

	θ_2	$\mathbb{E}T_0 = \mathbb{E}T_H$	$\mathbb{E}T_G$
A	$1/2$	$5/9$	$5/9$
B	$5/9$	$5/9$	$115/189$

. (C.3)

These follow from $q(1/2, \log 2) = 2/3$ and $q(5/9, \log 2) = 5/7$. Their target rates differ by $1/18$. For arbitrary $B > 0$, the small-bias settings in Appendix D.2 give the same risk classification.

C.3 A COUNT-DISTANCE BOUND USED IN THE PROOFS

We use the classical orthogonal expansion for sums of independent Bernoulli variables (Zacharovas, 2019).

Lemma C.1 (Distance from the matching binomial law)**.** *Let $X_i \sim \mathrm{Ber}(p_i)$ be independent, $i = 1, \dots, N$, with $N \geq 2$. Put $p = N^{-1}\sum_i p_i \in (0,1)$, $a_i = (p_i - p)/\sqrt{p(1-p)}$, and $\zeta = N^{-1}\sum_i a_i^2$. For $P = \mathcal{L}(\sum_i X_i)$ and $Q = \mathrm{Bin}(N, p)$,*

$$\chi^2(P\|Q) = \sum_{j=2}^{N} \frac{e_j(a)^2}{\binom{N}{j}}, \tag{C.4}$$

where e_j is the jth elementary symmetric polynomial. If $e\zeta < 1$, then

$$\frac{N}{2(N-1)}\zeta^2 \leq \chi^2(P\|Q) \leq \frac{(e\zeta)^2}{1-e\zeta}. \tag{C.5}$$

Proof. Let $P_0 = \bigotimes_{i=1}^N \mathrm{Ber}(p_i)$ be the law of $X = (X_1, \dots, X_N)$, and let $Q_0 = \mathrm{Ber}(p)^{\otimes N}$ be the product reference law. Let $Z_i = (X_i - p)/\sqrt{p(1-p)}$. The products $Z_I = \prod_{i\in I} Z_i$ form an orthonormal family. The product likelihood ratio is $dP_0/dQ_0 = \prod_i (1 + a_i Z_i)$. For a deterministic observation $S = T(X)$, let P_S, Q_S be its laws under P_0, Q_0. Its likelihood ratio is

$$\frac{dP_S}{dQ_S}(S) = \mathbb{E}_{Q_0}\left[\frac{dP_0}{dQ_0}(X) \,\middle|\, S\right],$$

as follows by testing against any bounded function of S.

For $S = \sum_i X_i$, define $B_j(S) = \binom{N}{j}^{-1/2} \sum_{|I|=j} Z_I$. This symmetric function depends only on S, and $B_0, \dots, B_N$ are orthonormal under Q. Exchangeability gives $\mathbb{E}_{Q_0}[Z_I \mid S] = B_j(S)/\sqrt{\binom{N}{j}}$ for $|I| = j$. Projecting the product likelihood onto S and applying Parseval's identity gives (C.4); the $j = 1$ term vanishes because $\sum_i a_i = 0$.

Write $A_2 = \sum_i a_i^2 = N\zeta$. For complex z,

$$\left|\prod_i (1 + a_i z)\right|^2 \leq \exp\left(2\,\mathrm{Re}(z)\sum_i a_i + |z|^2 A_2\right) = e^{A_2|z|^2}.$$

Cauchy's coefficient bound on the circle of radius $\sqrt{j/A_2}$ gives $|e_j(a)| \leq (eA_2/j)^{j/2}$ when $A_2 > 0$. Since $\binom{N}{j} \geq (N/j)^j$, the jth term is at most $(e\zeta)^j$; summing proves the upper bound. Finally $e_2 = -A_2/2$, so the second term alone proves the lower bound. The case $A_2 = 0$ is immediate. □

C.4 PROOF OF COROLLARY 1

The membership verification above and Theorem 2 give the claims for O, $O \cup H$, and $O \cup G$. To prove the remaining claim for G alone for every $B > 0$, choose a fixed sufficiently small $h > 0$ and set

$$b(h) = \mathrm{logit}(1/2 + 2h) - \mathrm{logit}(1/2 - h) \leq B.$$

Compare $(\theta_1, \theta_2, a_\star, b) = (1/2, 1/2 - h, 3, b(h))$ and $(1/2, 1/2, 1, b(h))$. The respective count laws of G are

$$P_{h,m} = \mathcal{L}\{\mathrm{Bin}(m, 1/2 + 2h) + \mathrm{Bin}(2m, 1/2 - h)\}, \qquad Q_m = \mathrm{Bin}(3m, 1/2). \tag{C.6}$$

The two binomial terms in $P_{h,m}$ are independent. Their common mean is $3m/2$. In Lemma C.1, $\zeta = 8h^2$, so $\chi^2(P_{h,m}\|Q_m) \leq Ch^4$ uniformly in m. Choose h small enough that total variation is at most $1/2$. Equation (B.11) gives a positive lower bound independent of m. The constant estimator gives the matching upper bound. Student training is postprocessing, while the rate for $O \cup G$ observed through students follows from Theorem 2.

C.5 JOINT RECOVERY UNDER A COMMON PRIOR

For a quantitative version with $B \geq \log 2$, add $\eta_C = (1/2, 5/9, 1, \log 2)$ and let π be the uniform prior on $\{\eta_A, \eta_B, \eta_C\}$. For a collection X, let $R^*_{\pi,X}$ be the minimum π-average mean squared error for θ_2 over estimators using the complete observations of X. The two laws

$$p_m = \mathcal{L}\{\mathrm{Bin}(2m, 1/2) + \mathrm{Bin}(m, 2/3)\}, \qquad q_m = \mathrm{Bin}(3m, 5/9) \tag{C.7}$$

describe the only differing count in the A/B subproblem for O and the A/C subproblem for G. For O, the reference observations have an identical independent law under A, B. Let $\Delta = 1/18$. Dropping the third nonnegative risk term and minimizing pointwise over an estimator gives, for $X = O, G$,

$$R^*_{\pi,X} \geq \frac{\Delta^2}{3} \sum_j \frac{p_m(j) q_m(j)}{p_m(j) + q_m(j)}. \tag{C.8}$$

The identity used here is $\inf_z \{p(z-u)^2 + q(z-v)^2\} = pq(u-v)^2/(p+q)$.

For this count pair, $\zeta = 1/40$. Set $D_* = (e/40)^2/(1 - e/40)$ and $v_* = \sqrt{D_*}/2 < 0.03520$. Lemma C.1 and Cauchy–Schwarz imply $\mathrm{TV}(p_m, q_m) \leq v_*$ for all m. Since $pq/(p+q) \geq \min(p,q)/2$,

$$R^*_{\pi,O}, R^*_{\pi,G} \geq \frac{\Delta^2}{6}(1 - v_*). \tag{C.9}$$

The square root of this bound exceeds 2.2277 percentage points.

For joint observations, the pairs $(\mathbb{E}T_{R_2}, \mathbb{E}T_G)$ are $(5/9, 5/9)$, $(5/9, 115/189)$, and $(1/2, 5/9)$. Their minimum ℓ_∞ separation is $\gamma = 10/189$. Selecting the closest pair and returning its target rate is correct whenever both coordinates deviate from their means by less than $\gamma/2$. Hoeffding's inequality and a union bound therefore give

$$R^*_{\pi,O\cup G} \leq 4\Delta^2 \exp(-3m\gamma^2/2). \tag{C.10}$$

The joint observations thus give exponential convergence under this three-point prior.

D PROOFS FOR RETRAINING AND UNBIASED LABELS

Throughout this section, R batches use independent source data and synthetic randomness conditional on the common unknown parameter. The additional-label policy may adaptively choose tasks, with at most k labels over the entire procedure. Write $r_{m,n} = (m^{-1} + n^{-1})^{-1}$ and $r_{m,\infty} = m$.

D.1 A MOMENT ESTIMATOR FROM INDEPENDENT TRAINING BATCHES

Consider one target model with $N = Jm$ training labels. At most one constituent source has probability $\theta + d$, and the others have probability θ. Put $\mu = \theta + d/J$ and $\delta = d/J \geq 0$. Summing Bernoulli variances gives

$$v_S := \frac{\mathrm{Var}\, S}{N} = \mu(1-\mu) - (J-1)\delta^2, \qquad \theta = \mu - \sqrt{\frac{\mu(1-\mu) - v_S}{J-1}}. \tag{D.1}$$

The formula also holds with $d = 0$ when the biased source is absent.

For $R \geq 2$, let

$$\widehat{\mu} = \frac{1}{NR} \sum_{r=1}^{R} S_r, \qquad \widehat{v} = \frac{1}{N(R-1)} \sum_{r=1}^{R} (S_r - \overline{S})^2, \tag{D.2}$$

and estimate

$$\widehat{\theta} = \Pi_I \left[\widehat{\mu} - \sqrt{\frac{[\widehat{\mu}(1-\widehat{\mu}) - \widehat{v}]_+}{J-1}} \right]. \tag{D.3}$$

For $\sigma_S^2 = \mathrm{Var}\, S$ and $\mu_4 = \mathbb{E}(S - \mathbb{E}S)^4$, independence gives

$$\mu_4 = 3\sigma_S^4 + \sum_{\ell=1}^{N} p_\ell(1-p_\ell)[1 - 6p_\ell(1-p_\ell)] \leq \frac{3N^2}{16} + \frac{N}{4} \leq \frac{5N^2}{16}. \tag{D.4}$$

The usual sample-variance expansion then yields

$$\mathrm{Var}(\widehat{v}) = \frac{1}{N^2 R}\left(\mu_4 - \frac{R-3}{R-1}\sigma_S^4\right) \le \frac{3}{8R}, \qquad \mathbb{E}(\widehat{\mu}-\mu)^2 \le \frac{1}{4NR}. \tag{D.5}$$

For $R = 2$ use $\mu_4 + \sigma_S^4 \le 3N^2/8$; for $R \ge 3$ discard the nonpositive term. Since $u(1-u)$ is 1-Lipschitz on $[0,1]$ and $(\sqrt{[x]_+} - \sqrt{y})^2 \le |x-y|$ for $y \ge 0$, projection and Cauchy–Schwarz give

$$\mathbb{E}(\widehat{\theta}-\theta)^2 \le \frac{1}{2NR} + \frac{1}{J-1}\left(\frac{1}{\sqrt{NR}} + \sqrt{\frac{3}{2R}}\right) \lesssim R^{-1/2}. \tag{D.6}$$

For $R = 1$ return $1/2$. Apply this estimator to T_0 for O, and to the candidate itself for G. Alternatively, $k \ge 1$ unbiased target-task labels give their empirical mean with risk at most $1/(4k)$. Choosing between these estimators by the known budgets gives risk $O((\sqrt{R}+k)^{-1})$.

D.2 Matching lower bounds for O and G

Set $J = 3$ and

$$h_0 = \min\{1/(8J), 3B/(16J)\}, \qquad b(h) = \mathrm{logit}(1/2 + (J-1)h) - \mathrm{logit}(1/2 - h). \tag{D.7}$$

For $0 < h \le h_0$, all probabilities lie in I and $b(h) \le (16/3)Jh \le B$. For O, compare

$$\eta_A(h) = (1/2, 1/2 - h, 3, b(h)), \qquad \eta_B(h) = (1/2, 1/2, 5, b(h)). \tag{D.8}$$

The reference laws agree, and their independent target count laws are exactly (C.6). For G alone, replace source 5 in $\eta_B(h)$ by source 1. This gives the same target count comparison.

Lemma C.1 has $\zeta = 4(J-1)h^2$ here. Consequently, uniformly over $m \ge 1$,

$$\mathrm{KL}(P_{h,m} \| Q_m) \le \chi^2(P_{h,m} \| Q_m) \le C_J h^4, \qquad C_J = 512(J-1)^2. \tag{D.9}$$

KL adds over independent batches. Conditional on any common interaction history, the policy uses the same action kernel in both settings. An unbiased reference-task label has the same law, while each unbiased target-task label contributes at most $4h^2$ to KL by (B.12). Padding stopped paths with a deterministic symbol and applying the chain rule gives

$$\mathrm{KL}(\mathsf{P}_A \| \mathsf{P}_B) \le C_J R h^4 + 4k h^2. \tag{D.10}$$

Take $h^2 = c/(\sqrt{R}+k)$, where $c = \min\{h_0^2, 1/64, 1/(8\sqrt{C_J})\}$. The right-hand side is at most $5/64 < 1/8$. Pinsker and (B.11) yield

$$\mathcal{R}_X(m, \infty; R, k) \ge \frac{c}{16(\sqrt{R}+k)}, \qquad X \in \{O, G\}. \tag{D.11}$$

Together with (D.6), this proves the first two complete-observation entries of Table 1, with constants independent of m, R, k.

D.3 Fast rates with an additional model

For any fixed configuration satisfying $\mathcal{U}(C) = \varnothing$, average each transformed observation V_i over the R batches. Its variance is divided by R. Applying (B.8) to this average gives risk at most $C/(R r_{m,n})$. Choosing between this estimator and the mean of k target labels gives

$$\mathcal{R}_{\mathcal{C}}(m, n; R, k) \lesssim (R r_{m,n} + k)^{-1}. \tag{D.12}$$

For the lower bound, use the unbiased submodel and reveal all source labels, as in Appendix B.3. There are at most $ARm + k$ independent target labels, giving $\Omega((Rm+k)^{-1})$. For student observations, revealing synthetic sampling positions instead exposes at most $L_2 Rn + k$ distinct target labels, giving $\Omega((Rn+k)^{-1})$. The adaptive-label chain-rule argument above remains valid in these enlarged experiments. Since $r_{m,n} \asymp \min\{m, n\}$, the two lower bounds imply

$$\mathcal{R}_{\mathcal{C}}(m, n; R, k) \asymp (R r_{m,n} + k)^{-1}. \tag{D.13}$$

Under complete observations only the source-label bound is needed. Applying this result to $O \cup G$ proves its complete-observation entry in Table 1.

D.4 One synthetic label per model

For the configurations in (D.8), the target count has mean $N/2$ in both cases. The expected smoothed teacher probability is therefore $(N/2+1)/(N+2) = 1/2$. A single target synthetic label has the same $\mathrm{Ber}(1/2)$ law. For O, the reference count vector has the same joint law and is independent of the target block, so the entire student observation law is identical. The argument for G alone uses its corresponding two settings from Appendix D.2. Independent repetition preserves the equality for every R.

Only the additional labels contribute KL, at most $4kh^2$. Taking $h^2 = c_1/(1+k)$ with $c_1 \le \min\{h_0^2, 1/32\}$ proves an $\Omega((1+k)^{-1})$ lower bound. Target labels, or the constant estimate when $k = 0$, attain that order. For $O \cup G$, (D.13) and $r_{m,1} = m/(m+1) \in [1/2, 1)$ give $\Theta((R+k)^{-1})$. This completes Theorem 3.

D.5 One target model with an ambiguous pair

The two slow entries of Table 1 extend beyond Figure 1.

Proposition D.1 (Retraining with one target model)**.** *In the family of Section 3.3, with any fixed $B > 0$, suppose the collection contains exactly one target model and $\mathcal{U}(C) \neq \varnothing$. Then, for all $m, R \ge 1$ and $k \ge 0$,*

$$\mathcal{R}_{\mathcal{C}}(m, \infty; R, k) \asymp (\sqrt{R} + k)^{-1}, \qquad \mathcal{R}_{\mathcal{C}}(m, 1; R, k) \asymp (1+k)^{-1}. \tag{D.14}$$

Proof. The moment estimator of Appendix D.1, applied to the target model, and the mean of k unbiased target-task labels give the first upper bound; with one synthetic label per model, the labels or the constant estimate $1/2$ give the second. For the lower bounds, fix $(a, c) \in \mathcal{U}(C)$; the target model contains a and omits c. Replace sources 3 and 5 in (D.8) by a and c. Since a and c have the same reference pattern, the reference laws agree, and the target count laws are

$$P_{h,m} = \mathcal{L}\big\{\mathrm{Bin}(m, 1/2+(J-1)h) + \mathrm{Bin}((J-1)m, 1/2-h)\big\}, \qquad Q_m = \mathrm{Bin}(Jm, 1/2). \tag{D.15}$$

The KL bound of Appendix D.2 and the single-label argument of Appendix D.4 then apply without change. □

A universal slow law for every configuration with an ambiguous pair would be false. Take sources a, b, c, d, reference models $\{a, c\}$ and $\{b, d\}$, and target models $\{a, b\}$ and $\{a, d\}$, so $J = 2$ and $(a, c) \in \mathcal{U}(C)$. At $m = 1$, the target counts are $S_1 = Y_a + Y_b$ and $S_2 = Y_a + Y_d$ with independent Bernoulli labels. Let $Z = \mathbf{1}\{|S_1 - S_2| = 1\}$ and $T = Z \min\{S_1, S_2\}$. On $\{Z = 1\}$, exactly one of Y_b, Y_d is positive, so $\min\{S_1, S_2\} = Y_a$, and independence gives

$$q_a = \frac{\mathbb{E}T}{\mathbb{E}Z}, \qquad q_b = \mathbb{E}S_1 - q_a, \qquad q_d = \mathbb{E}S_2 - q_a. \tag{D.16}$$

Retained probabilities are uniformly interior, so $\mathbb{E}Z$ is bounded below. At most one source is biased, and only upward, so $\theta_2 = \min\{q_a, q_b, q_d\}$. Plugging in moments averaged over R batches, with the denominator floored at its lower bound, gives risk $O(R^{-1})$; revealing all source labels in the unbiased subfamily gives the matching $\Omega(R^{-1})$, and k labels give $\Theta((R+k)^{-1})$. Several target models can thus expose source information through their joint discrete law, beyond the mean-and-variance mechanism of Section 4. This does not contradict the one-batch lower bound as $m \to \infty$ in Theorem 2.

D.6 The same two settings under four observation schemes

Let $H^2(P, Q) = \int(\sqrt{dP} - \sqrt{dQ})^2 \in [0, 2]$. For the pair (D.8), Table D.1 makes the distributional mechanism explicit.

For any statistic T and constant c_0, Cauchy–Schwarz gives

$$|\mathbb{E}_P T - \mathbb{E}_Q T|^2 \le 2H^2(P, Q)\{\mathbb{E}_P(T - c_0)^2 + \mathbb{E}_Q(T - c_0)^2\}. \tag{D.17}$$

For O, the independent reference factor is common. Set $T = (S - N/2)^2/N$ and $c_0 = 0$. The target laws have expectations differing by $2h^2$, while the fourth-moment bound above gives a uniform bound on $\mathbb{E}_P T^2 + \mathbb{E}_Q T^2$. Thus $H^2 \gtrsim h^4$; the reverse bound follows from $H^2 \le \chi^2$ and Lemma C.1.

Table D.1: **Squared Hellinger distances for one training batch.** Constants are uniform in $m \geq 1$ and $0 < h \leq h_0$. The comparison uses the same pair (D.8) throughout.

Collection	Observations	Squared distance
O	Complete observations	$\Theta(h^4)$
$O \cup G$	Complete observations	$\Theta(\min\{1, mh^2\})$
O	One synthetic label per model	0
$O \cup G$	One synthetic label per model	$\Theta(h^2)$

For G in this same pair, the difference in expected empirical positive-label rates is $d_G(h) = \{q(1/2, b(h)) - 1/2\}/3 \asymp h$. Taking $T = S_G$ and c_0 to be the midpoint of its two means in (D.17), and using variance at most $N/4$, yields

$$H^2(P_{S_G}, Q_{S_G}) \geq \frac{N^2 d_G(h)^2}{N + N^2 d_G(h)^2} \gtrsim \min\{1, mh^2\}.$$

The joint observation retains this marginal information. Revealing all source labels gives the reverse bound: their probabilities differ by $O(h)$ and there are $14m$ labels, so KL is $O(mh^2)$.

The zero-distance entry follows from Appendix D.4. For $O \cup G$ with one synthetic label per model, the marginal success probabilities for G differ by $Nd_G(h)/(N+2) \asymp h$, giving a lower bound of order h^2. For the upper bound, reveal the sampled positions as in Appendix B.3. At most L distinct source labels are exposed; each contributes $O(h^2)$ to KL, uniformly in m. This proves all four entries.

D.7 The label budget under sparse source bias

We prove Theorem 4 for the single-task family of Section 2. The observations consist of $S = CK$, equivalently the complete fitted probabilities, and k labels $Z_\ell \overset{\text{iid}}{\sim} \text{Ber}(\theta)$ independent of K. Observing all k labels also covers policies that stop earlier, since their observations are a randomized transformation of this experiment.

Proof of Theorem 4. Upper bounds and the sample-size lower bound. When $\kappa_s(W) > 0$, let $\widehat{\theta}_m$ be the estimator in Appendix G.1, whose risk is at most C_0/m. Independence and the zero mean of $Z_\ell - \theta$ give

$$\mathbb{E}\left(\frac{m\widehat{\theta}_m + \sum_{\ell=1}^k Z_\ell}{m+k} - \theta\right)^2 \leq \frac{C_0 m + k/4}{(m+k)^2} \lesssim (m+k)^{-1}. \tag{D.18}$$

In the unbiased subfamily $b = 0$, revealing all source labels and the additional labels gives $Am + k$ independent Bernoulli observations. Two rates separated by $c/\sqrt{Am+k}$, for a sufficiently small fixed $c > 0$, have total variation distance at most $1/2$. The two-point bound (B.11) yields $\Omega((Am + k)^{-1}) = \Omega((m+k)^{-1})$. For $\kappa_s = 0$, the mean of the additional labels has risk at most $1/(4k)$ when $k \geq 1$; use $1/2$ when $k = 0$. This gives the upper bound $O((1+k)^{-1})$.

A continuous family of ambiguous settings. Suppose $\kappa_s(W) = 0$. Choose $v \in \mathbb{R}^A$ with $Wv = \mathbf{1}$ and support T of size at most $2s$. Partition $T = T_0 \dot{\cup} T_1$ with $|T_j| \leq s$, and let $v^{(j)}$ retain the coordinates of v in T_j, with zero elsewhere. For $h \geq 0$, set

$$\begin{aligned} \theta^{(0)}(h) &= \tfrac{1}{2}, & e^{(0)}(h) &= hv^{(0)}, & q^{(0)}(h) &= \tfrac{1}{2}\mathbf{1} + hv^{(0)}, \\ \theta^{(1)}(h) &= \tfrac{1}{2} + h, & e^{(1)}(h) &= -hv^{(1)}, & q^{(1)}(h) &= \tfrac{1}{2}\mathbf{1} + h(\mathbf{1} - v^{(1)}). \end{aligned} \tag{D.19}$$

Each setting has at most s nonzero source effects. At $h = 0$ both are unbiased. By continuity, a fixed $\bar{h} > 0$ can be chosen so that, for every $h \in [0, \bar{h}]$, the rates lie in I, all source probabilities lie in $[1/4, 3/4]$, and the induced shifts $b_a^{(j)}(h) = \text{logit}\, q_a^{(j)}(h) - \text{logit}\, \theta^{(j)}(h)$ have magnitude at most B. Since $W\mathbf{1} = \mathbf{1}$,

$$W\big(q^{(0)}(h) - q^{(1)}(h)\big) = hW(v - \mathbf{1}) = 0.$$

Thus the complete count vectors have identical means for every h.

Overlap uniform in the label budget. Write the columns of C in a basis of their integer lattice as $w_a \in \mathbb{Z}^d$. They span $\mathbb{R}^d$, so $G_0 = \sum_a w_a w_a^\top \succ 0$. Put $u^{(0)} = v^{(0)}$ and $u^{(1)} = \mathbf{1} - v^{(1)}$. In these coordinates the one-layer count covariances are exactly

$$\Sigma_j(h) = \tfrac{1}{4}G_0 - h^2 \sum_a (u_a^{(j)})^2 w_a w_a^\top, \qquad j \in \{0,1\}. \tag{D.20}$$

Both converge to $G_0/4$ as $h \to 0$ and remain uniformly positive definite. Shrink $\bar{h}$ so that the total variation distance between their centered Gaussian laws is at most $1/8$ for all $h \le \bar{h}$. Lemma B.2 applies uniformly over the compact probability family. For the complete count laws $P_{j,h,m}$ it therefore gives a single m_0, independent of h and k, such that

$$\sup_{0 \le h \le \bar{h}} \mathrm{TV}(P_{0,h,m}, P_{1,h,m}) \le \tfrac{1}{4}, \qquad m \ge m_0. \tag{D.21}$$

The additional-label laws are $Q_{0,h,k} = \mathrm{Ber}(1/2)^{\otimes k}$ and $Q_{1,h,k} = \mathrm{Ber}(1/2+h)^{\otimes k}$. Taking $\bar{h} \le 1/4$ ensures $\mathrm{KL}(Q_{0,h,k} \| Q_{1,h,k}) \le 8kh^2$. The product structure and Pinsker's inequality yield

$$\mathrm{TV}(P_{0,h,m} \otimes Q_{0,h,k}, P_{1,h,m} \otimes Q_{1,h,k}) \le \tfrac{1}{4} + 2\sqrt{k}\, h. \tag{D.22}$$

Choose $h = c_*/\sqrt{1+k}$, where $c_* = \min\{\bar{h}, 1/16\} > 0$. This distance is at most $3/8$. Since the population rates differ by h, (B.11) gives

$$\mathcal{R}_C(m, \infty; 1, k) \ge \frac{c_*^2}{16(1+k)}, \qquad m \ge m_0, \quad k \ge 0. \tag{D.23}$$

The uniformity in (D.21) allows k to grow arbitrarily with m while leaving m_0 unchanged. □

D.8 Additional labels with shared cross-task selection

For the two-task family of Section 3.3, the counterpart of Theorem 4 is

$$\mathcal{R}_{\mathcal{C}}(m, \infty; 1, k) \asymp \begin{cases} (m+k)^{-1}, & \mathcal{U}(C) = \varnothing, \\ (1+k)^{-1}, & \mathcal{U}(C) \ne \varnothing, \quad m \ge m_0. \end{cases} \tag{D.24}$$

The threshold m_0 is independent of k. The case $\mathcal{U}(C) = \varnothing$ follows from (D.13) with $R = 1$.

Otherwise fix $(a, c) \in \mathcal{U}(C)$. For small $h \in [0, \bar{h}]$, choose $b(h)$ so that $d(1/2, b(h)) = Jh$ and compare

$$\eta_A(h) = (1/2, 1/2, a, b(h)), \qquad \eta_B(h) = (1/2, 1/2 + h, c, b(h)).$$

Here $b(h) = \mathrm{logit}(1/2 + Jh)$, and $\bar{h} > 0$ can be fixed small enough for admissibility. Reference laws agree; target observation means are all $1/2 + h$. In the target observation's lattice coordinates, let w_j be the source columns, $G_0 = \sum_j w_j w_j^\top \succ 0$, and $v = w_a$, while $w_c = 0$. The one-layer covariances are

$$\Sigma_A(h) = \tfrac{1}{4}G_0 - J^2 h^2 v v^\top, \qquad \Sigma_B(h) = (\tfrac{1}{4} - h^2) G_0. \tag{D.25}$$

They coincide at zero and remain uniformly positive definite for small $\bar{h}$. Shrink $\bar{h}$ so their Gaussian total variation distances are at most $1/8$. The uniform approximation in Lemma B.2 gives an m_0 such that

$$\sup_{0 \le h \le \bar{h}} \mathrm{TV}(P_{A,h,m}, P_{B,h,m}) \le 1/4, \qquad m \ge m_0. \tag{D.26}$$

For any additional-label policy and fixed initial observation o, reference labels have the same law and each target label contributes at most $8h^2$ to conditional KL. If $K_A(o, \cdot), K_B(o, \cdot)$ are the resulting history kernels, the chain rule gives $\sup_o \mathrm{KL}(K_A(o,\cdot) \| K_B(o,\cdot)) \le 8kh^2$. Insert an intermediate experiment with the same initial law and use total variation contraction:

$$\mathrm{TV}(P_A K_A, P_B K_B) \le \mathrm{TV}(P_A, P_B) + \sup_o \mathrm{TV}(K_A(o,\cdot), K_B(o,\cdot)) \le \tfrac{1}{4} + 2\sqrt{k}\, h. \tag{D.27}$$

Set $h^2 = \min\{\bar{h}^2, [256(k+1)]^{-1}\}$. The distance is at most $3/8$, and (B.11) gives

$$\mathcal{R}_{\mathcal{C}}(m, \infty; 1, k) \ge \frac{\min\{\bar{h}^2, 1/256\}}{16(k+1)}, \qquad m \ge m_0. \tag{D.28}$$

The bound transfers to student observations by simulating student training from the complete models. For $k \ge 1$ the target-label mean has risk at most $1/(4k)$; for $k = 0$ use $1/2$. This proves the second case of (D.24) and its counterpart for student observations. Uniformity in (D.26) is what makes m_0 independent of the label budget.

D.9 Students trained on at least two synthetic labels

Consider one target model in the scalar nonnegative single-source-bias family. It uses $J \geq 2$ sources per batch, with $N = Jm$, pooled count S, and fitted probability $P = (S+1)/(N+2)$. For $R \geq 2$ independently retrained batches, observe $Z_r \mid P_r \sim \mathrm{Bin}(n, P_r)$ with $n \geq 2$. Write $U_r = Z_r/n$, $\mu_P = \mathbb{E}P$, and $v_P = \mathrm{Var}(P)$. The factorial moments satisfy

$$\mathbb{E}\frac{(Z)_j}{(n)_j} = \mathbb{E}P^j \quad (j \leq n), \qquad \mathrm{Var}(U) = \left(1 - \frac{1}{n}\right) v_P + \frac{\mu_P(1-\mu_P)}{n}. \tag{D.29}$$

Here $(z)_j = z(z-1)\cdots(z-j+1)$.

Use the sample mean $\bar{U}$ and unbiased sample variance $\widehat{v}_U$ across batches to define

$$\widehat{v}_P = \frac{\widehat{v}_U - \bar{U}(1-\bar{U})/n}{1 - 1/n}, \qquad \widehat{\mu} = \mathrm{clip}_{[0,1]} \frac{(N+2)\bar{U} - 1}{N}, \qquad \widehat{v}_S = \frac{(N+2)^2}{N}\widehat{v}_P. \tag{D.30}$$

The corrected estimate is

$$\widehat{\theta}_2 = \mathrm{clip}_{[1/4,3/4]} \left[\widehat{\mu} - \sqrt{\frac{[\widehat{\mu}(1-\widehat{\mu}) - \widehat{v}_S]_+}{J-1}} \right]. \tag{D.31}$$

Proposition D.2 (An attainable risk bound for student observations)**.** *For fixed J and the parameter bounds above, uniformly over $m \geq 1$, $n \geq 2$, and $R \geq 2$,*

$$\mathbb{E}(\widehat{\theta}_2 - \theta_2)^2 \leq C \min\{1, (1 + m/n)/\sqrt{R}\}. \tag{D.32}$$

Proof. Decompose $U - \mu_P = (P - \mu_P) + (U - P)$. The fourth moments of the two terms are bounded by Cm^{-2} and Cn^{-2}, respectively. Thus $\mathbb{E}(U-\mu_P)^4 \leq C(m^{-1}+n^{-1})^2$. The sample variance has expected absolute error at most $C(m^{-1}+n^{-1})/\sqrt{R}$, while $\mathbb{E}(\bar{U}-\mu_P)^2 \leq C(m^{-1}+n^{-1})/R$.

The factor $(1-1/n)^{-1}$ is at most two. Since $u(1-u)$ is 1-Lipschitz on $[0,1]$, plugging $\bar{U}$ into (D.29) and multiplying by $(N+2)^2/N = O(m)$ gives

$$\mathbb{E}|\widehat{v}_S - v_S| \leq \frac{C(1+m/n)}{\sqrt{R}}, \qquad v_S = \mathrm{Var}(S)/N.$$

The corrected mean has error of smaller order. Equation (11) gives $\theta_2 = \mu - \sqrt{[\mu(1-\mu) - v_S]/(J-1)}$. Applying $|\sqrt{a_+} - \sqrt{b}|^2 \leq |a-b|$ for $b \geq 0$ and then clipping proves the claimed bound; bounded loss supplies the constant cap. □

E Experimental Methods and Additional Results

The matched-candidate and controlled LoRA experiments use the memberships in Table C.1; the acquisition and conditional-prediction experiments use the configurations specified below. RMSE is reported in percentage points. Conditional-prediction Brier scores and differences are multiplied by 10^3; rate MSE differences are multiplied by 10^4.

E.1 Generalized least-squares estimation

For each possible column $c \in \{0,1\}^L$ of C for the biased source, we fit (θ_1, θ_2, b) to the transformed observations V_i from Appendix A.2. Following (7), with the column $C_{:,a_\star}$ replaced by c, the candidate mean vector $\mu_c(\theta_1, \theta_2, b) \in \mathbb{R}^L$ has entries

$$[\mu_c(\theta_1, \theta_2, b)]_i = \theta_{t_i} + \frac{c_i}{J}\left[q(\theta_{t_i}, b) - \theta_{t_i}\right]. \tag{E.1}$$

The objective is

$$Q_c(\theta_1, \theta_2, b) = [V - \mu_c(\theta_1, \theta_2, b)]^\top \Omega [V - \mu_c(\theta_1, \theta_2, b)], \qquad \Omega = \Gamma^\dagger, \tag{E.2}$$

where $\Gamma^\dagger$ is the Moore–Penrose inverse of the working covariance

$$\Gamma_{ij} = \frac{C_i C_j^\top \mathbf{1}\{t_i = t_j\}}{4J^2 m} + \mathbf{1}\{i = j\}\frac{\alpha_m^2}{4n}. \tag{E.3}$$

The second term is zero under complete observations. When calibration labels are available, let k_t be the number drawn for task t, with $k_1 + k_2 \le k$. For each task with $k_t > 0$, the empirical rate is appended as an observation whose mean θ_t has no source-bias term, with working variance $1/(4k_t)$ and zero working covariance with the model observations.

We enumerate distinct source columns and minimize Q_c by bounded Gauss–Newton updates. The bounds are $b \in [0, 1]$ and $\theta_t \in [1/4, 3/4]$ for the analytic simulation, or $\theta_t \in [0.02, 0.98]$ for the controlled LoRA experiment. Numerically tied minima are averaged, weighted by source-column multiplicity. GLS optimization settings are shared across methods within each study.

The pairwise-overlap baseline receives the same $O \cup G$ model values and calibration labels, together with the total source count, sources per model, and all pairwise overlaps. It fits the same objective over all possible membership columns for the biased source consistent with these constraints, averaging tied minima with equal weights.

E.2 Analytic simulations

We draw independent source counts under settings A and B of Appendix C.2, reuse them according to the source records, and apply the training rules in Appendix A.2. Experiments observing fitted probabilities directly use $m \in \{32, 128, 512, 2048, 8192\}$; experiments observing students fix $m = 2048$ and vary $n \in \{128, 512, 2048, 8192, 32768\}$. There are 2,048 replicates per setting at each sample size. Methods share source counts and synthetic draws wherever their acquired models coincide.

For replicate r, setting w, and method a, the target-task loss is

$$\ell_{r,w,a} = (\widehat{\theta}_{r,w,a,2} - \theta_{w,2})^2, \qquad \widehat{R}_a = \frac{1}{2} \sum_{w \in \{A,B\}} \frac{1}{2048} \sum_{r=1}^{2048} \ell_{r,w,a}. \tag{E.4}$$

Figure E.1 plots $100\sqrt{\widehat{R}_a}$. We compute normal-approximation 95% intervals for the equally weighted mean loss across settings, then take square roots of the nonnegative endpoints.

The Bayes reference for $O \cup H$ knows both settings and their equal prior probabilities. It computes the joint target-count likelihood by summing over the shared source count and returns the posterior mean of θ_2. The reference-task likelihood is identical under both settings and cancels.

Table E.1 also reports the existing collection alone and acquisition of both candidates.

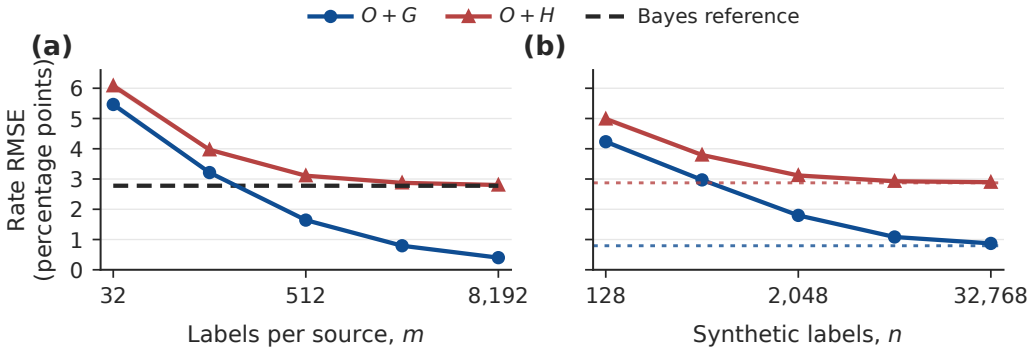

Figure E.1: **Per-source and synthetic sample budgets.** (a) Rate RMSE with complete model observations. (b) Synthetic-label delivery at $m = 2{,}048$; dotted lines give complete-observation errors. Both panels weight settings A and B equally. Shading gives 95% intervals from 2,048 replicates per setting.

Table E.1: **Rate RMSE under complete model observations.** Errors are in percentage points, with equal weights for settings A and B. The Bayes reference uses $O + H$ and the known two-setting prior.

m	O	$O + H$	$O + G$	$O + G + H$	Bayes reference
32	6.55	6.08	5.46	5.03	2.78
128	4.13	3.97	3.21	3.06	2.78
512	3.15	3.11	1.64	1.56	2.78
2,048	2.89	2.87	0.79	0.75	2.78
8,192	2.80	2.80	0.40	0.37	2.78

E.2.1 Adding unbiased labels to a fixed batch of models

This simulation uses the single-task model of Section 2 with $s = 2$, $B = 1$, and the nine-source configuration in Table H.1. In setting 1, $\theta = 0.45$, sources 7 and 9 have retained rate 0.54, and all others have rate 0.45. In setting 2, $\theta = 0.48$, source 3 has rate 0.57, source 4 has rate 0.39, and all others have rate 0.48. The observation means for O_9 coincide across settings, while those for G_9 differ.

For each setting and $m \in \{128, 512, 2048, 8192\}$, we draw 1,024 independent training batches. Each batch is reused for O_9 at every label budget $k \in \{0, 32, 128, 512, 2048, 8192, 32768\}$ and for $O_9 \cup G_9$ at $k = 0$. Additional labels are independent $\mathrm{Ber}(\theta)$ draws, nested as k increases. Model observations are the complete fitted probabilities, with smoothing undone as in Section 2.

For each acquired collection, the estimator enumerates source pairs and fits $\vartheta\mathbf{1} + Wf$ by generalized least squares, restricting f to that pair and using working covariance $WW^\top/(4m)$. With $k > 0$, the additional-label mean is appended with mean ϑ, variance $1/(4k)$, and zero covariance with model observations. We choose the minimum-residual fit and clip its rate to I. Fits use source records and observations, without the biased-source identities or the two parameter settings. Figure 2a averages the two settings' MSEs before taking the square root; Figure E.2 displays them separately. These risks illustrate the resource comparison of Theorem 4 at two fixed settings.

Table E.2 pools the two settings at every training budget. On this grid, unbiased labels first reach the accuracy of $O_9 \cup G_9$ at $k = m$ when $m = 128$ and at $k = 4m$ for every larger m; at $k = m$ they leave 1.3–2.1 times its RMSE for $m \geq 512$.

Table E.2: **Pooled rate RMSE with unbiased labels.** Squared errors from the two settings are averaged before taking the square root, as in Figure 2a. $O_9 \cup G_9$ uses no additional labels. Monte Carlo standard errors are at most 0.04 points.

m	$O_9 \cup G_9$	O_9, $k = 0$	O_9, $k = m$	O_9, $k = 4m$
128	2.72	3.07	2.56	1.88
512	1.32	2.47	1.73	1.06
2,048	0.46	2.14	0.97	0.43
8,192	0.22	2.09	0.29	0.18

E.3 Model acquisition with two biased sources

We use 12 pre-specified configurations, each with seven models combining three of nine possible sources on one target task. Ten have $\kappa_2 = 0$; the other two already satisfy $\kappa_2 > 0$. For each of the ten deficient configurations, candidates are all previously unused three-source combinations drawn from sources present in the existing collection. Structural selection maximizes the augmented margin (4), breaking ties by source-index order. Every configuration admits a candidate with positive margin. The comparator samples uniformly from the same candidate set, independently of model observations. Both acquisition methods then fit by least squares over the possible sets of biased sources, as in Theorem 1.

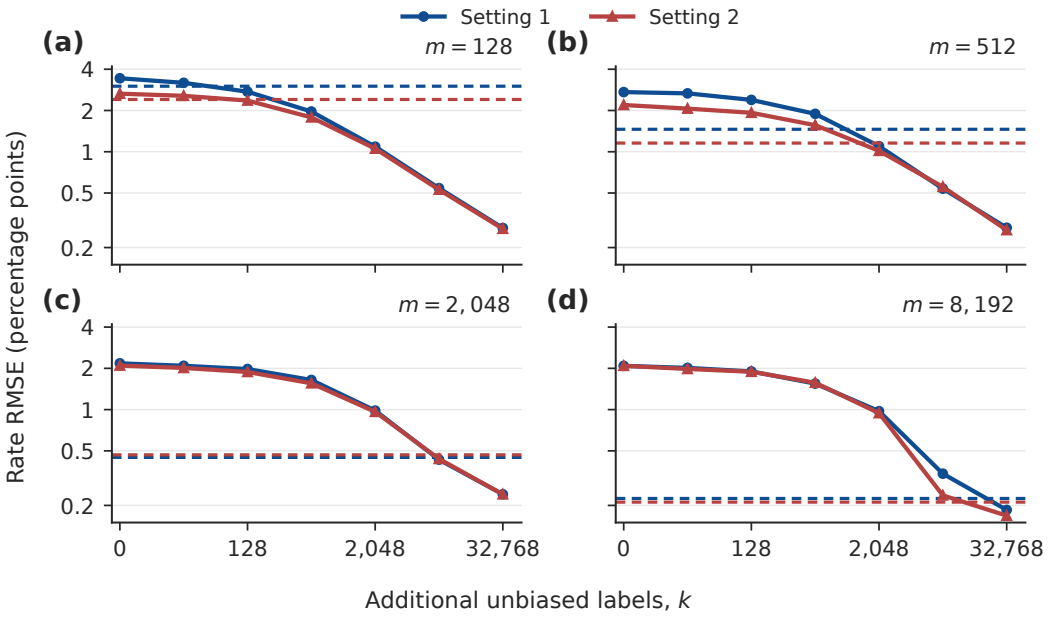

Figure E.2: **Unbiased labels and model acquisition at four training budgets.** Solid curves show O_9 supplemented by k unbiased labels; dashed lines show $O_9 \cup G_9$ without additional labels. Each setting has 1,024 replicates. Within a replicate, all label budgets share the same model observations, and additional labels are nested across budgets.

For each configuration we generate six new parameter settings: $\theta \in [0.35, 0.65]$, two randomly located biased sources, and bias magnitudes in $[0.3, 0.9]$. Three settings use positive biases and three use opposite signs. At each $m \in \{64, 256, 1024, 4096, 16384\}$, we use $n = 4m$ and 1,024 sampling replicates per setting. Within each replicate, methods share the source counts and existing model observations; identical acquisitions also share synthetic labels. Acquisition uses only the source records and the bound $s = 2$.

We average MSE equally across the ten configurations and six settings, then report its square root in percentage points. Table E.3 gives the aggregate results and the number of settings in which structural selection has lower empirical MSE than random acquisition. The paired MSE improvement at the largest sample size is 0.37 squared percentage points, with a Monte Carlo 95% interval of $[0.35, 0.39]$ conditional on these configurations and parameters.

Table E.3: **Acquisition using source records.** RMSE is in percentage points. The last column counts parameter settings favoring structural over random acquisition.

m	Existing	Random +1	Structural +1	Settings
64	5.49	5.24	5.15	34/60
256	3.25	2.99	2.83	36/60
1,024	1.65	1.38	1.12	44/60
4,096	1.04	0.79	0.42	36/60
16,384	0.89	0.64	0.20	30/60

E.4 Acquiring models trained on human feedback

Data and source selection. We use the general-domain HelpSteer3 feedback subset (Wang et al., 2025), with a positive label when a majority of annotators rate the response mostly or perfectly helpful. The fixed labeled pool contains 3,000 labeled contexts. In each replicate, nine sources receive disjoint blocks of $m = 128$ retained records. Sources 7 and 9 undergo selection; the four pre-

specified conditions combine label-dependent or input-dependent retention with a positive or negative direction. Let z_1 be the first standardized input feature. The log-odds shift is $b_a(x) = \pm 1.5$ for label-dependent selection and $b_a(x) = \pm[1 + \sigma(z_1)]$ for input-dependent selection, with a common sign at the two affected sources. Labels 1 and 0 are retained with probabilities $\exp[\min(b_a(x), 0)]$ and $\exp[-\max(b_a(x), 0)]$, respectively. A fifth condition applies no selection.

We use 512 public unlabeled texts for distillation and a separate panel of 128 unlabeled inputs for selecting the correction. The calibration pool contains 256 contexts, from which each replicate draws 128 labels for the labeled controls. Evaluation uses 1,024 new contexts. The public correction inputs, calibration pool, and evaluation set are disjoint from one another and from development and distillation contexts. Model, correction, and comparison settings were fixed before evaluation.

Training and delivery. Frozen `Qwen3-Embedding-0.6B` features (Zhang et al., 2025) are projected to 32 dimensions using PCA and scaling fitted on earlier public inputs. Each source fits a regularized logistic predictor to its retained human labels. Each row of C in Table H.1 averages the predictions of its three source predictors. A separate network with one hidden layer of 32 GELU units and a sigmoid output learns each mixture by binary cross-entropy on the 512 distillation inputs. Networks share initialization and training settings within each replicate and use 400 optimization steps. Their training targets are the selected-source mixture probabilities.

All networks are saved before acquisition and reloaded for downstream prediction. The learner receives these networks, source records, and public inputs. Source predictors and training records are not delivered. The five selection conditions and 16 paired replicates produce 720 networks; methods share the same trained networks whenever their acquired collections overlap. Averaged over delivered networks, the function RMSE relative to the source mixture is 0.43 percentage points.

Acquisition and correction. The membership criterion selects G_9 from the two candidates before their outputs are inspected. For each acquired collection, write $W = C/3$ and use the common-support estimator in Appendix H.2 with $s = 2$ and weighting matrix $(WW^\top)^{-1/2}$. For each source pair, fit the target probability and source effects to the observed predictions. The pair with the smallest mean weighted residual on the 128 public correction inputs defines one coefficient vector used at all subsequent inputs. Thus the same estimation rule applies to O_9, $O_9 \cup H_9$, $O_9 \cup G_9$, and $O_9 \cup G_9 \cup H_9$.

Raw predictions, ordinary averaging, and source correction are computed and fixed before any additional calibration labels are used. Labeled controls then receive the same 128 labels. Platt scaling fits a logistic regression to the logit of the average model probability; output stacking fits all acquired probabilities, with a logit-input version also reported below. Label-based source selection chooses among the candidate source-pair fits by calibration Brier loss. We also report a smoothed label mean and the adaptive combination of ordinary and corrected predictions from Appendix H.3.

Evaluation. Let $\mathcal{C}_b$ contain the four biased conditions, and let $\widehat{f}_{r,c,a}$ denote the prediction from replicate r, condition c, and method a. On the $N = 1024$ evaluation records, with $\bar{y} = N^{-1}\sum_j y_j$, we compute

$$\widehat{B}_a = \frac{1}{16 \cdot 4 \cdot N} \sum_{r=1}^{16} \sum_{c \in \mathcal{C}_b} \sum_{j=1}^{N} (\widehat{f}_{r,c,a}(x_j) - y_j)^2, \tag{E.5}$$

$$\mathrm{RMSE}_{a,\mathrm{rate}} = 100 \sqrt{\frac{1}{16 \cdot 4} \sum_{r=1}^{16} \sum_{c \in \mathcal{C}_b} \left(\frac{1}{N} \sum_{j=1}^{N} \widehat{f}_{r,c,a}(x_j) - \bar{y} \right)^2}. \tag{E.6}$$

The three primary Brier comparisons evaluate $O_9 \cup G_9$ against $O_9 \cup H_9$ and O_9 under the common correction rule, and the raw G_9 against raw H_9 predictions. We use 5,000 paired crossed bootstrap draws over replicate and evaluation-context indices, preserving methods and conditions within each draw. Intervals have nominal coverage $1 - 0.05/3$ for a Bonferroni family error level of 0.05. Secondary comparisons use descriptive 95% intervals. Sampling replicates share the fixed labeled pool and configuration.

Table E.4: **Primary comparisons for learned-model acquisition.** Brier differences are multiplied by 10^3 and subtract the second predictor's loss from the first. Paired bootstrap intervals have nominal coverage $1 - 0.05/3$ (approximately 98.33%).

Comparison	Difference	Interval
Joint correction: $O_9 + H_9$ versus $O_9 + G_9$	8.75	$[6.09, 11.23]$
Acquisition: O_9 versus $O_9 + G_9$	8.11	$[5.35, 10.65]$
Individual prediction: G_9 versus H_9	33.07	$[30.40, 36.11]$

Table E.5: **Acquisition and additional-label controls.** Brier scores are multiplied by 10^3; rate RMSE is in percentage points. Values average the four biased conditions and 16 paired replicates. The two groups use zero and 128 additional calibration labels, respectively.

Predictor	Brier	Rate RMSE	Labels
Raw G_9	279.78	21.01	0
Raw H_9	246.72	10.33	0
Source correction, O_9	243.55	9.41	0
Source correction, $O_9 + H_9$	244.19	9.71	0
Source correction, $O_9 + G_9$	235.44	2.52	0
Source correction, $O_9 + G_9 + H_9$	235.05	1.37	0
Prediction averaging, $O_9 + G_9$	245.69	10.50	0
Platt scaling, $O_9 + G_9$	236.52	3.41	128
Platt scaling, $O_9 + H_9$	236.58	3.40	128
Output stacking, $O_9 + G_9$	239.08	3.57	128
Logit stacking, $O_9 + G_9$	239.63	3.58	128
Label-based source selection, $O_9 + H_9$	236.55	4.07	128
Calibration labels alone	236.33	3.51	128
Adaptive correction, $O_9 + G_9$	235.81	3.23	128

Table E.6: **Rate recovery by selection condition.** RMSE is in percentage points. Individual columns use raw predictions; joint columns use the common source correction rule.

Selection	Raw G_9	Raw H_9	O_9	$O_9 + H_9$	$O_9 + G_9$
Label-dependent, positive	16.61	8.33	8.33	8.37	3.20
Label-dependent, negative	24.82	12.11	10.44	11.41	1.55
Input-dependent, positive	16.55	8.29	8.33	8.35	3.21
Input-dependent, negative	24.48	11.91	10.31	10.36	1.55
None	2.59	1.83	2.11	1.96	1.67

Table E.4 reports the primary comparisons. Table E.5 separates controls by their additional-label budgets, and Table E.6 gives the condition-specific rate errors, including no selection. We decompose Brier loss into squared error of the average prediction relative to the observed positive-label rate, and the remaining error after centering predictions and labels by their respective evaluation averages. In this decomposition, the $O_9 \cup H_9$ to $O_9 \cup G_9$ Brier improvement of 8.75 consists of 8.80 from reduced error in the average prediction and -0.05 from the error after centering, all scaled by 10^3. With 128 additional labels, Platt scaling reduces the $O_9 \cup H_9$ rate RMSE to 3.40 points; the Brier difference between this predictor and $O_9 \cup G_9$ source correction has a descriptive interval containing zero. The source-correction gain over output stacking has a positive descriptive interval.

E.5 Adaptive calibration and transfer to neural students

This separate experiment studies calibration strength and transfer of corrected predictions to students. Its two configurations and selection conditions are specified below.

Data and representations. We use the general-domain HelpSteer3 feedback subset (Wang et al., 2025). A response is positive when a majority of the available annotators describe it as mostly or perfectly helpful. The labeled pool contains 3,000 labeled contexts. A separate panel contains 256 calibration contexts, 512 unlabeled distillation contexts, and 1,024 evaluation contexts. The panel is disjoint from development data, and model and correction settings were fixed before evaluation. Each replicate draws 128 calibration labels from the calibration pool. Frozen `Qwen3-Embedding-0.6B` features (Zhang et al., 2025) are projected to 32 dimensions by a PCA and scaling transform fitted on earlier public unlabeled inputs.

Sources and observed predictors. The five-source configuration contains all mixtures omitting one source. The nine-source configuration uses the seven rows of the first pre-specified acquisition configuration and its selected eighth row (Appendix E.3). Every predictor averages the source predictions specified by its row of W. Within a replicate, sources receive disjoint retained blocks of $m = 256$ human-labeled records, and models sharing a source reuse its fitted predictor. Each source fits a regularized logistic head on the frozen features. The correction method observes mixed probabilities, source records, and public inputs.

Sources 1 and A undergo selection, where A is the number of sources. Positives are retained with probability one and negatives with probability $\exp[-b_a(x)]$. The label-dependent setting uses strengths 1.5 and 1.8. In the input-dependent setting, with standardized features z, the two strengths are $0.6 + 1.8\sigma(z_1 + 0.3z_2)$ and $0.6 + 1.8\sigma(-z_1 + 0.3z_2)$. A third setting applies no selection. These rules operate only in constructing source training blocks; the correction algorithm uses the bound $s = 2$.

Correction, calibration, and students. For each candidate pair of biased sources, we compute the weighted least-squares residual on all 512 public inputs with $H = (WW^\top)^{-1/2}$. The pair with smallest average residual supplies one fixed coefficient vector for predicting at every input, as in Appendix H.2. The adaptive predictor minimizes calibration Brier loss over convex combinations of ordinary averaging and full source correction using (H.11). Logistic recalibration fits the ordinary average probability as its input score; output stacking fits a logistic model to all mixed probabilities. Both use all 128 calibration labels. Additional controls fit a logistic classifier to public features, or jointly to features and model outputs, using the same labels.

Ordinary and corrected students have one hidden layer of 32 GELU units and a sigmoid output on the frozen features. They share initialization and are trained by binary cross-entropy against their respective soft predictions on the 512 distillation inputs. The source heads and students are fitted anew within each sampling replicate. Training sources, calibration labels, and evaluation contexts are paired across methods.

Evaluation and uncertainty. Let $\widehat{f}_{r,g,c,a}$ be the predictor for replicate r, configuration g, selection condition c, and method a. With $N = 1024$ evaluation records and $\mathcal{C}_b$ the two biased conditions, the

primary aggregate loss is

$$\widehat{B}_a = \frac{1}{16 \cdot 2 \cdot 2 \cdot N} \sum_{r=1}^{16} \sum_{g=1}^{2} \sum_{c \in \mathcal{C}_b} \sum_{j=1}^{N} (\widehat{f}_{r,g,c,a}(x_j) - y_j)^2. \tag{E.7}$$

The two pre-specified differences compare ordinary averaging with adaptive correction, and ordinary with adaptive students. Intervals use 5,000 paired crossed bootstrap draws: replicate indices and evaluation-context indices are resampled independently, preserving methods, configurations, and conditions within each draw. The 97.5% percentile intervals implement Bonferroni adjustment for the two primary comparisons. Secondary comparisons use descriptive 95% intervals. All are conditional on the fixed labeled pool and configurations.

Table E.7: **Paired gains on human feedback.** Brier improvements are multiplied by 10^3. The primary intervals are Bonferroni-adjusted; secondary intervals are descriptive.

Comparison	Gain	Interval	Confidence
Correction vs. averaging	4.98	$[2.38, 7.76]$	97.5%, primary
Student vs. ordinary student	4.94	$[2.35, 7.72]$	97.5%, primary
Correction vs. recalibration	2.09	$[1.16, 3.25]$	95%, secondary
Correction vs. stacking	3.05	$[1.68, 4.68]$	95%, secondary

Table E.8: **Human-feedback prediction under equal calibration budgets.** Brier scores are multiplied by 10^3. Each column averages both configurations; the biased column also averages label-dependent and input-dependent selection.

Method	Selection bias	No selection
Prediction averaging	240.99	234.84
Full source correction	235.01	235.38
Adaptive correction	236.01	235.10
Progressive aggregation	236.22	234.92
Logistic recalibration	238.10	237.75
Output stacking	239.06	239.11
Ordinary student	240.98	234.88
Adaptive student	236.04	235.14
Input-feature classifier	253.49	253.49
Input-feature and output stacking	254.01	254.01

Table E.9: **Correction gains by configuration and selection condition.** Brier differences, multiplied by 10^3, compare prediction averaging with adaptive correction. Intervals are descriptive 95% paired bootstrap intervals.

Layout	Selection	Improvement	Interval
Five sources	None	-0.21	$[-0.80, 0.29]$
Five sources	Label-dependent	10.43	$[5.94, 15.10]$
Five sources	Input-dependent	8.99	$[4.91, 13.31]$
Nine sources	None	-0.32	$[-0.98, 0.13]$
Nine sources	Label-dependent	0.27	$[-0.10, 0.74]$
Nine sources	Input-dependent	0.23	$[-0.09, 0.66]$

Table E.7 gives the paired gains and intervals for adaptive correction and its students. Table E.8 reports biased and unselected performance under the same calibration budget; Table E.9 separates the configurations and selection conditions. In a descriptive decomposition, 4.91 of the 4.98 scaled Brier improvement corresponds to reduced error in the average prediction over evaluation inputs; the error after centering contributes 0.07. The principal empirical benefit is therefore probability calibration. Student predictions approximate their corresponding teacher functions with RMSE of 0.17–0.18 percentage points on evaluation inputs.

E.6 NEURAL STUDENTS WITH SIMULATED FEEDBACK

Checkpoint and texts. We use `Skywork-Reward-V2-Qwen3-0.6B` (Liu et al., 2026) and HelpSteer2 prompt–response texts (Wang et al., 2024), without their human ratings. Disjoint prompt groups provide 288 reference texts, 64 distillation texts, and 512 evaluation texts per task, with normalized duplicates removed between evaluation and development pools. Task assignments and text pools remain fixed across the 16 feedback replicates. Frozen reward scores $g(x)$ use normalization fixed during development.

Feedback and probability recovery. For each task, let $\mathcal{X}_t^{\mathrm{ref}}$ be its reference texts. Define

$$F_t(v, x) = \sigma(g(x) + a_t(v)), \qquad \frac{1}{|\mathcal{X}_t^{\mathrm{ref}}|} \sum_{x \in \mathcal{X}_t^{\mathrm{ref}}} F_t(v, x) = v. \tag{E.8}$$

The unique offset is found by scalar root solving. For each source–task pair we sample reference inputs with replacement, draw feedback with probability $F_t(\theta_t, x)$, and apply selection until $m = 4096$ labels are retained. The biased source keeps all positives and independently keeps each negative with probability $1/2$. Source blocks are then shared according to Table C.1. The settings are A and B from Appendix C.2.

Each replicate also generates 128 unbiased calibration labels per task from the reference inputs, shared by all four methods. GLS uses the calibration rates and model counts to estimate $\widehat{\theta}_t$, yielding soft training probabilities $F_t(\widehat{\theta}_t, x)$. Common random streams pair the settings and methods.

Student training. Students use the same frozen base checkpoint with rank-8 LoRA adapters (Hu et al., 2022) and a task-specific affine calibration head. They minimize binary cross-entropy with the soft targets on the 128 distillation texts. All methods share initialization, text order, and optimization settings within each replicate.

Evaluation and intervals. For the $M = 512$ target-task evaluation texts, write

$$e_{r,w,a} = \frac{1}{M} \sum_{i=1}^{M} \{\widehat{p}_{r,w,a}(x_i) - F_2(\theta_{w,2}, x_i)\}, \qquad \ell_{r,a} = \frac{1}{2} \sum_{w \in \{A,B\}} e_{r,w,a}^2. \tag{E.9}$$

Student RMSE is $100\sqrt{16^{-1} \sum_r \ell_{r,a}}$. Soft-target RMSE replaces $\widehat{p}$ by $F_2(\widehat{\theta}_{r,w,a,2}, x)$; transfer RMSE uses the difference between mean student and mean soft-target probabilities. Each setting contributes a squared error before averaging. Student rate RMSE is 3.44 points for O, 3.38 for $O \cup H$, 0.73 for $O \cup G$, and 2.37 for the pairwise-overlap comparator. Soft-target RMSE for $O \cup G$ is 0.53 points. Across methods, student-to-soft-target mean-probability RMSE is 0.44–0.50 points.

Table E.10 reports differences $\ell_{r,\mathrm{control}} - \ell_{r,O \cup G}$ with paired t intervals across the 16 replicates: 97.5% for the two pre-specified comparisons under Bonferroni adjustment, and a descriptive 95% interval for the overlap comparator.

Table E.10: **Paired neural-student comparisons.** Positive differences favor source-membership correction with $O \cup G$.

Control	MSE difference	Interval	Coverage
O	11.27	$[5.97, 16.58]$	97.5%, primary
$O \cup H$	10.90	$[5.64, 16.17]$	97.5%, primary
Pairwise overlaps, $O \cup G$	5.11	$[0.19, 10.02]$	95%, secondary

F EXTENSIONS

F.1 KNOWN MONOTONE SELECTION CHANNELS

Replace the logistic channel by known task-specific functions $q_t(p, b)$, continuously differentiable in a neighborhood of the compact parameter domain, with

$$q_t(p,0) = p, \quad \epsilon_q \le q_t(p,b) \le 1-\epsilon_q, \quad 0 < c_q \le \partial_b q_t \le C_q, \quad \partial_p q_t \ge 0 \text{ or } |\partial_p(q_t - p)| \le D < J. \tag{F.1}$$

Retained labels are independent Bernoulli draws with these probabilities, and the same source parameter b links the two tasks. The training and delivery rules are unchanged.

Proposition F.1. *Under* (F.1)*, Theorem 2 and condition* (9) *remain valid.*

Proof. The proof of Lemma B.1 uses only the derivative bounds in (F.1), reference coverage, and $q_t(p, 0) = p$; under the second alternative, take $a_0 = 1 - D/J$. Continuous differentiability on the compact domain bounds $|\partial_p(q_t - p)|$, as required for A_J. The resulting upper bound is therefore unchanged. The unbiased submodel gives the two sample-budget lower bounds. If $(a, c) \in \mathcal{U}(C)$, choose small $b_* > 0$ and $\Delta = [q_2(1/2, b_*) - 1/2]/J > 0$. The two settings in Appendix B.5 have identical reference observation distributions and equal target observation means. Interior probabilities and identical integer loadings allow Lemma B.2 to apply. The acquisition argument depends only on membership. □

F.2 UNEQUAL SOURCE COUNTS

Keep the family of Section 3.3, with a single nonnegative bias shared across tasks and any fixed $B > 0$, but let model i use its own known number $J_i \ge 1$ of sources, so that $W_{ia} = C_{ia}/J_i$. Write $W^{(t)}$ for the rows of task t and define

$$\mathcal{U}_W(C) = \left\{(a,c) : C^{(1)}_{:,a} = C^{(1)}_{:,c}, \quad W^{(2)}_{:,a} = \lambda \mathbf{1} \text{ for some } \lambda > 0, \quad W^{(2)}_{:,c} = \mathbf{0}\right\}. \tag{F.2}$$

With equal source counts, $\mathcal{U}_W(C) = \mathcal{U}(C)$.

Proposition F.2 (Unequal source counts)**.** $\mathcal{R}_{\mathcal{C}}(m, n) \asymp m^{-1} + n^{-1}$ *if* $\mathcal{U}_W(C) = \varnothing$*. Otherwise* $\mathcal{R}_{\mathcal{C}}(m, n) \asymp 1$ *for all sufficiently large* m*, including complete observations.*

Proof. Write $d(p, b) = q(p, b) - p$ and $J_{\max} = \max_i J_i$, and let $e_0 = \|\mu(\eta) - \mu(\eta')\|_\infty$ for parameters with biased-source choices a, c. As in Lemma B.1, the reference minima give $|\theta_1 - \theta'_1| \le e_0$.

If the reference columns of a, c differ, interchange the parameters if needed to find a reference row i with $(C_{ia}, C_{ic}) = (1, 0)$. Comparing μ_i with $\mu'_i = \theta'_1$ gives $d(\theta_1, b) \le 2J_{\max}e_0$. Every reference elevation under η is then at most $2J_{\max}e_0$, so the two reference ranges bound $d(\theta'_1, b') \le 2J_{\max}(J_{\max} + 1)e_0$. Hence $b + b' \le Ce_0$, and every target row gives $|\theta_2 - \theta'_2| \le e_0 + C_q(b + b') \le Ce_0$.

If the reference columns coincide, a reference row containing both sources gives $|d(\theta_1, b) - d(\theta'_1, b')| \le 2J_{\max}e_0$, hence $|b - b'| \le (2J_{\max} + D)e_0/c_q$. For a target row i, write $w = W_{ia}$ and $w' = W_{ic}$. If $w = w'$ in some row, the function $F_w(p, b) = (1 - w)p + wq(p, b)$ has $\partial_p F_w \ge a_1 := \min\{1, \inf \partial_p q\} > 0$ on the compact domain, which gives $a_1|\theta_2 - \theta'_2| \le e_0 + C_q|b - b'| \le Ce_0$. Otherwise every target row contains exactly one of a, c. If both target columns have a zero entry, the target minima give $|\theta_2 - \theta'_2| \le e_0$. In the remaining case, one source, say a, appears in every target row and the other in none. If its weights differ in two rows i, j, then $\mu'_i = \mu'_j = \theta'_2$, while $\mu_i - \mu_j = (W_{ia} - W_{ja})d(\theta_2, b)$. Thus

$$d(\theta_2, b) \le \frac{2e_0}{|W_{ia} - W_{ja}|}, \qquad |\theta_2 - \theta'_2| \le e_0 + d(\theta_2, b) \le Ce_0. \tag{F.3}$$

Constant weights are exactly the case $(a, c) \in \mathcal{U}_W(C)$. Hence, if $\mathcal{U}_W(C) = \varnothing$, the target rate is uniformly stable, and the minimum-residual fit of Appendix B.2 with the variance bound in (2) gives the upper bound. The two sample-budget lower bounds of Appendix B.3 use the unbiased subfamily and are unchanged.

For $(a, c) \in \mathcal{U}_W(C)$ with weight λ, fix a small $b_* > 0$ and compare

$$\eta_A = (1/2, 1/2, a, b_*), \qquad \eta_B = (1/2, 1/2 + \lambda d(1/2, b_*), c, b_*). \tag{F.4}$$

The two sources have the same reference pattern, so the reference count laws agree. Every target model contains a with weight λ and omits c, so all target observation means equal $1/2 + \lambda d(1/2, b_*)$ in both settings. The integer count loadings are the same, and the argument of Appendix B.5 gives a positive lower bound for all large m. Students inherit it by postprocessing. □

A nonempty $\mathcal{U}_W(C)$ requires $\lambda = 1/J_i$ for every target row, so all target models must use the same number of sources. If the existing target models all use J_0 sources, a target candidate using a set $\mathcal{S}$ resolves (a, c) exactly when $a \notin \mathcal{S}$, $c \in \mathcal{S}$, or $|\mathcal{S}| \neq J_0$: a different pooling weight is itself a comparison. Condition (9) is the case $|\mathcal{S}| = J_0$.

F.3 Several signed biases shared across tasks

Now allow up to s biased sources with either sign, shared across the two tasks: source a has retained rate $q(\theta_t, b_a)$ on task t, with $\theta_t \in I$, $\|b\|_0 \leq s$, and $|b_a| \leq B$ for any fixed $B > 0$. Models may use unequal known source counts. Let W_1 contain the reference rows and define

$$\alpha_s(W_1) = \min_{|T| \leq 2s} \sigma_{\min}\big([\mathbf{1}, (W_1)_{:,T}]\big), \qquad \sigma_{\min}(M) = \inf_{\|z\|_2 = 1} \|Mz\|_2. \tag{F.5}$$

Proposition F.3 (Several shared signed biases)**.** *If $\alpha_s(W_1) > 0$ and the collection contains a target model, then the minimax risk for θ_2 is $\Theta(m^{-1} + n^{-1})$, with $n^{-1} = 0$ for complete observations.*

Proof. All retained probabilities lie in a fixed compact subinterval of $(0, 1)$. Write $e_{1a} = q(\theta_1, b_a) - \theta_1$, so that the reference means are $\mu_1 = \theta_1 \mathbf{1} + W_1 e_1$. For two admissible parameters, the union of the bias supports has at most $2s$ elements, so

$$\big[(\theta_1 - \theta_1')^2 + \|e_1 - e_1'\|_2^2\big]^{1/2} \leq \alpha_s(W_1)^{-1} \|\mu_1 - \mu_1'\|_2. \tag{F.6}$$

Since $b_a = \mathrm{logit}(\theta_1 + e_{1a}) - \mathrm{logit}(\theta_1)$ and the logit is Lipschitz on the compact domain, $\|b - b'\|_2 \leq C\|\mu_1 - \mu_1'\|_2$.

For a target row with weights w_a, set $F(p, b) = \sum_a w_a q(p, b_a)$. On the compact domain, $\partial_p F \geq c_B > 0$, and $|\partial_{b_a} q| \leq 1/4$ gives $|F(p, b) - F(p, b')| \leq \|b - b'\|_2/4$. Hence

$$c_B|\theta_2 - \theta_2'| \leq |\mu_i - \mu_i'| + \tfrac{1}{4}\|b - b'\|_2 \leq C\|\mu - \mu'\|_2. \tag{F.7}$$

A minimum-residual fit of the complete mean vector is within twice the observation error of the true mean, and $\sum_i \mathrm{Var}(V_i) = O(m^{-1} + n^{-1})$ gives the upper bound. For the lower bounds, set $b = \mathbf{0}$ and fix θ_1; reference observations then carry no information about θ_2, and the arguments of Appendix B.3 apply. □

This condition is sufficient, not a classification: the reference models identify the shared effects, after which any target mixture is a known increasing function of θ_2. For every s, take $A = 2s + 1$ sources and A reference models that each omit one source. Then $W_1 = (\mathbf{1}\mathbf{1}^\top - I)/(A - 1)$ is invertible with $W_1^{-1}\mathbf{1} = \mathbf{1}$, which cannot be supported on $2s$ columns; hence $\alpha_s(W_1) > 0$, as in Appendix G.2. For instance, take $s = 2$, $\theta_1 = 2/5$, $\theta_2 = 3/5$, and $b = (\log 3, -\log 3, 0, 0, 0)$, so $B = \log 3 > 1$. The reference source rates are $(2/3, 2/11, 2/5, 2/5, 2/5)$, and the five reference means are $(19/55, 7/15, 68/165, 68/165, 68/165)$. A target model using sources $\{1, 2, 3\}$ has mean $289/495$, which the recovered effects invert to $\theta_2 = 3/5$. Reference observations alone carry no information about θ_2, and the target model alone cannot separate its rate from selection; jointly they give the parametric rate.

F.4 Approximate observation means and numerical error

Suppose $\mathcal{U}(C) = \varnothing$. Let the actual observation mean be μ_{act}, and suppose an admissible mean vector μ_0 with the same true population rates satisfies $\|\mu_{\mathrm{act}} - \mu_0\|_\infty \leq \delta$. Assume the coordinate variances obey (A.8). Let a fitted admissible mean $\widehat{\mu}$ satisfy

$$\|V - \widehat{\mu}\|_\infty \leq \inf_{\mu \text{ admissible}} \|V - \mu\|_\infty + \varepsilon_{\mathrm{opt}}.$$

Then, with $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$ as in Lemma B.1,

$$\mathbb{E}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_\infty^2 \le 3K^2 \left[L\left(\frac{1}{Jm} + \frac{\alpha_m^2}{n}\right) + 4\delta^2 + \varepsilon_{\mathrm{opt}}^2 \right]. \tag{F.8}$$

To prove this, write $e_V = \|V - \mu_{\mathrm{act}}\|_\infty$. Feasibility of μ_0 and the triangle inequality give $\|\widehat{\mu} - \mu_0\|_\infty \le 2e_V + 2\delta + \varepsilon_{\mathrm{opt}}$. Apply Lemma B.1, square using $(x+y+z)^2 \le 3(x^2+y^2+z^2)$, and bound $\mathbb{E}e_V^2$ by the sum of coordinate variances.

Fixed score strata provide one way to retain the selection equation. Freeze a stratum map $h_t(x)$ before fitting and define $\theta_{t,g} = \Pr(Y = 1 \mid t, h_t(X) = g)$ for each stratum g. If selection within each stratum depends on the label with the same source odds ratio, then

$$\Pr(Y = 1 \mid t, h_t(X) = g, \text{retained at } a) = q(\theta_{t,g}, b_a).$$

Fitting a separate probability model per stratum preserves the channel. Logistic channels on any fixed domain $[\epsilon_p, 1-\epsilon_p]$ satisfy (F.1) for every fixed B, since $\partial_p q > 0$; for $b \le 1$ one may take $c_q = \epsilon_p(1-\epsilon_p)/e$ and $C_q = 1/4$.

For a stratum-constant prediction and fixed population weights π_g, expansion of squared loss gives

$$R_{\mathrm{Brier}}(\widehat{\theta}) - R_{\mathrm{Brier}}(\theta) = \sum_g \pi_g (\widehat{\theta}_g - \theta_g)^2. \tag{F.9}$$

The cross term vanishes because $\mathbb{E}[Y - \theta_g \mid g] = 0$.

F.5 Acquiring several models

Let $\mathcal{C}$ denote the existing collection, and let a finite candidate library use the same sources and sampling mechanism, with each candidate using J sources. Its task and membership are known before acquisition. For a selected candidate set $\mathcal{T}$, $C_{\mathcal{C}\cup\mathcal{T}}$ holds the source records of the combined collection. Write $\mathfrak{U} = \mathcal{U}(C_{\mathcal{C}})$, and let candidate j use sources $\mathcal{S}_j$ on task t_j. Define the ambiguous source pairs resolved by that candidate as

$$D_j = \begin{cases} \{(a,c) \in \mathfrak{U} : \mathbf{1}\{a \in \mathcal{S}_j\} \ne \mathbf{1}\{c \in \mathcal{S}_j\}\}, & t_j = 1, \\ \{(a,c) \in \mathfrak{U} : a \notin \mathcal{S}_j \text{ or } c \in \mathcal{S}_j\}, & t_j = 2. \end{cases} \tag{F.10}$$

A reference candidate distinguishes reference membership patterns; a target candidate breaks the full-coverage/zero-coverage relation.

Proposition F.4 (Acquiring several candidates)**.** *For any selected candidate set $\mathcal{T}$,*

$$\mathcal{U}(C_{\mathcal{C}\cup\mathcal{T}}) = \mathfrak{U} \setminus \bigcup_{j\in\mathcal{T}} D_j. \tag{F.11}$$

The augmented collection permits consistent estimation exactly when the union covers $\mathfrak{U}$.

Proof. Adding reference rows can only separate previously equal reference columns. Adding target rows can only remove sources from the full-coverage or zero-coverage sets. Hence every pair that remains ambiguous was already in $\mathfrak{U}$. A pair survives precisely when all new reference rows contain both sources or neither, and every new target row contains a but omits c. This is equivalent to its absence from every D_j. Reference coverage remains valid after adding rows, so Theorem 2 applies to the augmented configuration. □

For specified positive acquisition costs c_j, the minimum cost of enabling consistent estimation is therefore the weighted set-cover problem

$$\min_{x_j \in \{0,1\}} \sum_j c_j x_j \quad \text{subject to} \quad \sum_{j:(a,c)\in D_j} x_j \ge 1 \quad \text{for all } (a,c) \in \mathfrak{U}. \tag{F.12}$$

The empty collection suffices if $\mathfrak{U}$ is empty; the problem is infeasible if the library cannot cover it. Otherwise, choosing the smallest cost per newly resolved pair gives the classical approximation ratio $\sum_{r=1}^{|\mathfrak{U}|} 1/r \le 1 + \log|\mathfrak{U}|$ (Chvátal, 1979). To see the bound, when r ambiguous pairs remain, some set in an optimal cover has cost per newly resolved pair at most OPT/r. Charging that ratio to each newly resolved pair and summing gives $\mathrm{OPT}\sum_{r=1}^{|\mathfrak{U}|} 1/r$.

G Recovery with Multiple Biased Sources and Imperfect Records

G.1 Sparse source bias without a cross-task link

Fix one target task, A sources, and L models with source records $C \in \{0,1\}^{L\times A}$. Row i contains $J_i \geq 1$ sources. Let $\theta \in I = [1/4, 3/4]$, $|b_a| \leq B$ for fixed $B > 0$, and $\|b\|_0 \leq s$, where $s \geq 1$. Positive and negative values of b_a allow selection in either direction. Retaining labels 0 and 1 with probabilities $\exp[-\max(b_a, 0)]$ and $\exp[\min(b_a, 0)]$, respectively, gives

$$q_a = \sigma(\operatorname{logit}\theta + b_a), \qquad K_a \overset{\text{ind}}{\sim} \operatorname{Bin}(m, q_a), \qquad S_i = \sum_a C_{ia}K_a, \qquad P_i = \frac{S_i + 1}{J_i m + 2}. \tag{G.1}$$

Models sharing a source reuse its count. Complete observations reveal S; students are trained on n independent synthetic labels per model, conditional on P. Writing their mean as U_i, set $V_i = (1 + 2/(J_i m))U_i - 1/(J_i m)$, or $V_i = S_i/(J_i m)$ under complete observations. With $W_{ia} = C_{ia}/J_i$ and $e_a = q_a - \theta$,

$$\mu = \mathbb{E}V = \theta\mathbf{1} + We, \quad \|e\|_0 \leq s, \qquad \sigma^2_{V,i} := \operatorname{Var}(V_i) \leq \frac{1}{4J_i m} + \frac{(1 + 2/(J_i m))^2}{4n}. \tag{G.2}$$

This decomposition is exact. It can be applied separately to tasks with different bias locations and strengths. Throughout this section, the configuration, s, and parameter bounds are fixed. Write $\mathcal{R}_C(m, n)$ for the minimax squared-error risk for θ, and use $n^{-1} = 0$ for complete observations.

Define the target separation margin

$$\kappa_s(W) = \min_{T\subseteq[A],\ |T|\leq 2s} \operatorname{dist}_2(\mathbf{1}, \operatorname{col}(W_{:,T}))\,, \tag{G.3}$$

where the column space of the empty matrix is $\{0\}$. This is a target-component version of classical sparse-error separation (Fawzi et al., 2014).

Proof of Theorem 1. For two settings, the difference $e - e'$ is supported on at most $2s$ sources. Projecting $\mu - \mu' = (\theta - \theta')\mathbf{1} + W(e - e')$ onto the orthogonal complement of those columns yields

$$|\theta - \theta'| \leq \kappa_s(W)^{-1}\|\mu - \mu'\|_2. \tag{G.4}$$

Enumerate supports of size at most s, minimize $\|V - \vartheta\mathbf{1} - Wf\|_2$ over $\vartheta \in \mathbb{R}$ and f on each support, and select a minimum-residual fit. Project its fitted ϑ onto I to obtain $\widehat{\theta}$. The observation mean μ is feasible, so the fitted vector is within $2\|V - \mu\|_2$ of μ. Projection onto I contracts the target error, giving

$$\mathbb{E}(\widehat{\theta} - \theta)^2 \leq 4\kappa_s(W)^{-2}\sum_i \sigma^2_{V,i}. \tag{G.5}$$

This proves the upper bound without knowledge of the biased sources or their selection functions.

For the sample-budget lower bounds, restrict to $b = 0$. Even observing all Am source labels gives risk $\Omega(m^{-1})$. For student observations, represent each synthetic label as a uniformly selected element of its model's $J_i m$ retained labels together with the fixed pseudolabels 0 and 1. Revealing the selected indices gives a more informative experiment. The indices are independent of θ, and the Ln draws reveal at most Ln distinct random labels. Thus its divergence between two population rates is at most $Ln\,\mathrm{kl}(\theta, \theta')$. Taking their separation of order $n^{-1/2}$ gives $\Omega(n^{-1})$. The larger of the two lower bounds controls their sum up to a constant.

If $\kappa_s(W) = 0$, choose T with $|T| \leq 2s$ and h satisfying $W_{:,T}h = \mathbf{1}$. Partition $T = T_0 \cup T_1$ with $|T_j| \leq s$. For an interior θ and a sufficiently small fixed $d > 0$, take $\theta' = \theta + d$, set $e = dh$ on T_0 and $e' = -dh$ on T_1, and set the other coordinates to zero. Then $e - e' = dh$ and the observation means coincide. For small enough d, both source-rate vectors are interior and the induced log-odds shifts have magnitude at most B.

Both count vectors are supported on the same lattice generated by the columns of C. In a lattice basis, the loadings are integer and span the full lattice, and their covariance matrices are positive definite. Lemma B.2 therefore applies: the total variation distance between the count laws converges to that between two nondegenerate Gaussian laws with the same mean, which is strictly below one. The two-point bound (B.11) gives a positive risk lower bound independent of m. Student observations are a common randomized transformation of the counts and inherit this bound. □

Unknown selection functions and approximate representativeness. Assume $\kappa_s(W) > 0$. The upper bound uses only independent source counts, the mixture weights, and sparsity of $q - \theta\mathbf{1}$. It therefore also holds when different sources use unknown selection mechanisms. More generally, suppose

$$q = \theta\mathbf{1} + e + u, \qquad \|e\|_0 \le s, \qquad \|u\|_\infty \le \delta, \tag{G.6}$$

with all source probabilities bounded away from zero and one. The same estimator satisfies

$$\mathbb{E}(\widehat{\theta} - \theta)^2 \le 4\kappa_s(W)^{-2} \left(\sum_i \sigma_{V,i}^2 + L\delta^2 \right). \tag{G.7}$$

Indeed, replace $V - \mu$ in the preceding proof by $\xi + Wu$, where ξ is centered sampling noise. Since $W\mathbf{1} = \mathbf{1}$, $\|Wu\|_\infty \le \delta$, and the expected cross term vanishes. The estimator requires no knowledge of δ. Over the family allowing all such small perturbations, the rate is $\Theta(m^{-1} + n^{-1} + \delta^2)$ for sufficiently small δ: take $\theta = p - \delta$, $u = \delta\mathbf{1}$, $\theta' = p + \delta$, $u' = -\delta\mathbf{1}$, and $e = e' = 0$. Both settings have identical source probabilities $p\mathbf{1}$, giving the δ^2 lower bound even with source-level observations. The unbiased subfamily supplies the other terms.

G.2 One additional model for any fixed number of biased sources

Let $A = 2s + 1$ and use all A mixtures that each omit one source: $C = \mathbf{1}\mathbf{1}^\top - I$ and $J = A - 1$. If μ_i is the mean of model i, then

$$q_i = \sum_{j=1}^{A} \mu_j - (A-1)\mu_i. \tag{G.8}$$

At least $s + 1$ source rates equal θ, so their median recovers θ stably. Equivalently, W is invertible and $W^{-1}\mathbf{1} = \mathbf{1}$, which cannot be supported on $2s$ columns; hence $\kappa_s(W) > 0$.

For the first $A-1$ models alone, source A appears in every model, so its column spans $\mathbf{1}$ and $\kappa_s = 0$. The last model alone also has zero separation margin. Theorem 1 thus gives nonvanishing risk for each collection separately and parameter-rate recovery jointly. Adding the last model reuses source blocks already present in the existing collection.

An explicit example has $A = 5$, $s = 2$, and $B = 1$:

$$\begin{aligned} \theta &= 2/5, \qquad q = (9/20, 2/5, 2/5, 2/5, 3/5), \\ \theta' &= 21/50, \quad q' = (47/100, 21/50, 21/50, 21/50, 27/50). \end{aligned} \tag{G.9}$$

Both settings have two positive, nonzero selection effects. The first four means coincide at $(9/20, 37/80, 37/80, 37/80)$, while the fifth means are $33/80$ and $173/400$. The first four count loadings generate $\mathbb{Z}^4$, so the complete-observation overlap argument applies directly.

G.3 Cross-task bias drift on the original configuration

Keep $O \cup G$ from Table C.1 and the single positively biased source, but allow reference and target log-odds shifts $b^{(1)}, b^{(2)} \in [0, B]$ with $|b^{(2)} - b^{(1)}| \le \varepsilon$; superscripts index tasks, while subscripts on b_a continue to index sources. Fix $B > 0$ and $0 \le \varepsilon \le \min\{B, 1\}/4$. Write $\mathcal{R}^{\mathrm{drift}}_{O\cup G}(m, n; \varepsilon)$ for the minimax risk for the target rate θ_2 over this enlarged family.

Theorem G.1 (Risk under cross-task drift)**.** *For all sufficiently large m, uniformly over the stated range of ε,*

$$\mathcal{R}^{\mathrm{drift}}_{O\cup G}(m, n; \varepsilon) \asymp m^{-1} + n^{-1} + \varepsilon^2. \tag{G.10}$$

The constants and the threshold for m may depend on B and the fixed configuration, but not on ε. The risk for $O \cup H$ remains of constant order.

Proof. Replace $b^{(2)}$ by $b^{(1)}$ while keeping the population rates fixed. The derivative of the logistic channel with respect to its log-odds shift is at most $1/4$, and a biased source has weight $1/3$. Thus every observation mean changes by at most $\varepsilon/12$. The stability bound (F.8) gives the upper bound.

For the drift lower bound, reveal that source 3 is biased, the reference rate is $1/2$, and $b^{(1)} = b_0 = B/2$. Define $F(p, b) = [2p + q(p, b)]/3$. Compare target parameters $(1/2, b_0)$ and $(p_\varepsilon, b_0 - \varepsilon)$, with

$$F(p_\varepsilon, b_0 - \varepsilon) = F(1/2, b_0). \tag{G.11}$$

The implicit derivative is

$$\frac{dp_\varepsilon}{d\varepsilon} = \frac{q(1-q)}{2 + q(1-q)/[p_\varepsilon(1-p_\varepsilon)]}. \tag{G.12}$$

On the relevant compact interior domain it is bounded below by a positive constant and above by $1/8$. Hence $c\varepsilon \le p_\varepsilon - 1/2 \le \varepsilon/8 \le 1/32$, so $p_\varepsilon \in I$, and $b_0 - \varepsilon \ge B/4$ keeps the shifted log-odds admissible. Both target models contain source 3, so their means coincide across these settings; the reference laws are identical. The target count loadings include the two coordinate vectors, generating $\mathbb{Z}^2$, and the covariance matrices are uniformly positive definite. Lemma B.2 gives uniformly positive distributional overlap and hence an $\Omega(\varepsilon^2)$ lower bound. The unbiased subfamily supplies the two sample-budget terms. For $O \cup H$, the original shared-bias family is contained in the drift family, retaining its nonvanishing lower bound. □

G.4 Incorrect source records

Let the reported matrix be $\widehat{C}$. Its true value differs in at most r rows, with each row's source count J_i correct and the source identities drawn from the same known set. These discrepancies are fixed before sampling or independent of the sampled labels. Let $\widehat{W}_{ia} = \widehat{C}_{ia}/J_i$ and define

$$\kappa_{s,r}(\widehat{W}) = \min_{D\subseteq[L],\ |D|\le 2r}\ \min_{T\subseteq[A],\ |T|\le 2s} \operatorname{dist}_2\Big(\mathbf{1}_{D^c}, \operatorname{col}(\widehat{W}_{D^c,T})\Big). \tag{G.13}$$

Theorem G.2 (Recovery with incorrect source records)**.** *Under the sparse-bias model* (G.1)*, $\kappa_{s,r}(\widehat{W}) > 0$ is sufficient for minimax risk $\Theta(m^{-1} + n^{-1})$ over the stated source-record uncertainty family.*

Proof. Enumerate deletions of at most r rows and bias supports of at most s sources. Minimize the Euclidean residual on the remaining rows and clip the fitted target to I. Deleting the actual erroneous rows makes the true observation mean feasible, with residual no greater than the norm of all sampling noise ξ. The fitted and actual deletion sets have union of size at most $2r$. On the jointly retained rows, the fitted vector and the true observation mean differ by at most $2\|\xi\|_2$. Projection as in (G.4) yields

$$\mathbb{E}(\widehat{\theta} - \theta)^2 \le 4\kappa_{s,r}(\widehat{W})^{-2} \sum_i \sigma_{V,i}^2. \tag{G.14}$$

The variance bounds hold under each possible true matrix. The correct-record, unbiased subfamily gives the matching sample-budget lower bounds. □

Feasible designs. For $s = 2, r = 1$, take any nine of the ten two-source mixtures on five sources. After deleting any two rows, the corresponding graph has at least seven edges and is connected and nonbipartite: a disconnected or bipartite simple graph on five vertices has at most six edges. Its unsigned edge–vertex incidence matrix has full column rank. Indeed, $Cx = 0$ implies $x_u = -x_v$ on each edge; an odd cycle forces zero values, which connectivity propagates. Since $W\mathbf{1} = \mathbf{1}$, the only representation of $\mathbf{1}$ uses all five columns, giving $\kappa_{2,1} > 0$.

For arbitrary fixed s, r, use all two-source mixtures on $A = 2\max(s, r+1) + 1$ sources. Deleting at most $2r$ edges from the complete graph leaves it connected because $A - 1 > 2r$, and nonbipartite because making it bipartite requires at least

$$\binom{A}{2} - \lfloor A^2/4 \rfloor = ((A-1)/2)^2 > 2r \tag{G.15}$$

edge deletions. The same rank argument, together with $A > 2s$, gives $\kappa_{s,r} > 0$. This redundancy uses additional combinations of existing source records.

An ambiguity caused by one uncertain source record. To show why recovery from each candidate matrix separately does not settle recovery with uncertain records, consider two five-row matrices. The rows of C_0 use source pairs $(1,2), (2,3), (3,4), (4,5), (5,1)$; C_1 changes only the last pair to $(1,3)$. Both matrices are invertible with determinant magnitude two, so knowing the matrix permits source-count inversion and median estimation at risk $O(m^{-1})$ when $s = 2$.

Set $B = 1$, $p = 2/5$, $t = 3/100$, and

$$\begin{aligned} \theta_0 &= p, & q_0 &= (p, p, p, p + 2t, p + 2t), \\ \theta_1 &= p + t, & q_1 &= (p + t, p - t, p + t, p + t, p + 3t). \end{aligned} \tag{G.16}$$

Each vector has two biased sources with log-odds shifts of magnitude below one, and $C_0 q_0/2 = C_1 q_1/2$. The C_0 count lattice requires the sum of all five counts to be even; the C_1 lattice requires the sum of counts 1, 2, and 5 to be even. Each has index two in $\mathbb{Z}^5$, and their intersection has index four. On a ball of radius proportional to $\sqrt{m}$ about the common mean, the local limit bound gives probability at least $cm^{-5/2}$ per intersection point under each law. There are at least $c'm^{5/2}$ such points. The laws therefore retain positive overlap, giving risk $\Omega(t^2)$ when the true matrix is unknown. The sufficient condition (G.13) supplies the redundancy that this example lacks.

H Conditional Prediction and Adaptive Correction

H.1 Prediction mixtures and function recovery

Let P be the population distribution of (X, Y) before selection, with $Y \in \{0, 1\}$ and $p(x) = \Pr(Y = 1 \mid X = x)$. If source a retains a record with probability $s_{a,y}(x) > 0$, its conditional positive-label rate is

$$q_a(x) = \frac{p(x) \exp[b_a(x)]}{1 - p(x) + p(x) \exp[b_a(x)]}, \qquad b_a(x) = \log \frac{s_{a,1}(x)}{s_{a,0}(x)}. \tag{H.1}$$

The functions b_a may vary with the input and are unknown to the estimator. At most s of the functions q_a differ from p. A source predictor $\widehat{q}_a$ is fitted to retained records. The fitted mixture and its population counterpart are

$$\widehat{f}_i(x) = \sum_a W_{ia} \widehat{q}_a(x), \qquad \mu_i(x) = \sum_a W_{ia} q_a(x), \qquad W_{ia} \geq 0, \quad W\mathbf{1} = \mathbf{1}. \tag{H.2}$$

The function actually delivered for model i is written $\widetilde{f}_i$. It equals $\widehat{f}_i$ when the mixture is delivered directly and equals a distilled student h_i otherwise. The downstream learner knows W and queries $\widetilde{f}_i$. This interface describes a composite prediction service. Source training records, source-model parameters, and selection logs are not delivered. With $e_a(x) = q_a(x) - p(x)$, the population mixtures satisfy $\mu(x) = p(x)\mathbf{1} + We(x)$ and $\|e(x)\|_0 \leq s$.

For any predictor f, the excess Brier risk is

$$\mathbb{E}_P(f(X) - Y)^2 - \mathbb{E}_P(p(X) - Y)^2 = \|f - p\|^2_{L^2(P_X)}. \tag{H.3}$$

Use the separation margin $\kappa_s(W)$ from (4).

Proposition H.1 (Stable conditional-probability recovery)**.** *Suppose $\kappa_s(W) > 0$. Given delivered functions $\widetilde{f}_i$, minimize $\|\widetilde{f}(x) - \vartheta\mathbf{1} - Wv\|_2$ over $\vartheta \in \mathbb{R}$ and $\|v\|_0 \leq s$ at each x, and clip the fitted ϑ to $[0, 1]$. The resulting function satisfies*

$$\mathbb{E}\|\widehat{p} - p\|^2_{L^2(P_X)} \leq 4\kappa_s(W)^{-2} \sum_i \mathbb{E}\|\widetilde{f}_i - \mu_i\|^2_{L^2(P_X)}. \tag{H.4}$$

Proof. At each input, the difference between the fitted and true source perturbations has support at most $2s$. Projecting onto the orthogonal complement of those columns gives the stability inequality (G.4). The population mixture $\mu(x)$ is feasible, so the fitted vector is within $2\|\widetilde{f}(x) - \mu(x)\|_2$ of it. Clipping contracts the target error. Square, integrate over P_X, and take expectation over training randomness. □

Errors from source predictors and distillation. Write $\epsilon_a(m) = \mathbb{E}\|\widehat{q}_a - q_a\|^2_{L^2(P_X)}$. Jensen's inequality gives

$$\mathbb{E}\|\widehat{f}_i - \mu_i\|^2_{L^2(P_X)} \leq \sum_a W_{ia} \epsilon_a(m). \tag{H.5}$$

Suppose model i is delivered as a student, $\widetilde{f}_i = h_i$, trained on N_{dist} distillation inputs from a distribution Q_i. Its population approximation error is $D_i(N_{\text{dist}}) = \mathbb{E}\|h_i - \widehat{f}_i\|^2_{L^2(Q_i)}$; no particular rate in N_{dist} is assumed. If $dP_X/dQ_i \le \rho_i$, then

$$\mathbb{E}\|\widehat{p} - p\|^2_{L^2(P_X)} \le \frac{8}{\kappa_s(W)^2} \sum_i \left[\rho_i D_i(N_{\text{dist}}) + \sum_a W_{ia} \epsilon_a(m) \right]. \tag{H.6}$$

This follows by inserting $h_i - \mu_i = (h_i - \widehat{f}_i) + (\widehat{f}_i - \mu_i)$ into (H.4). If the final corrected predictor is distilled directly into h, the corresponding bound is

$$\mathbb{E}\|h - p\|^2_{L^2(P_X)} \le 2\mathbb{E}\|\widehat{p} - p\|^2_{L^2(P_X)} + 2\rho D(N_{\text{dist}}), \qquad D(N_{\text{dist}}) = \mathbb{E}\|h - \widehat{p}\|^2_{L^2(Q)}, \tag{H.7}$$

where $dP_X/dQ \le \rho$. Thus errors in fitting source predictors, distillation, and input coverage enter separately. The quantities D_i, D are population function errors under the stated input distributions.

H.2 Selecting a shared biased-source support

A common source support can be selected using independent unlabeled inputs. Fix an invertible weighting matrix H, write $\mathbf{1}_H = H\mathbf{1}$ and $W_H = HW$, and define

$$\kappa_H = \min_{|T| \le 2s} \operatorname{dist}_2(\mathbf{1}_H, \operatorname{col}((W_H)_{:,T})). \tag{H.8}$$

For each support S of size s, let $\Phi_S = [\mathbf{1}_H, (W_H)_{:,S}]$, let Π_S be the orthogonal projector onto its column space, and define

$$\ell_S(x) = \|(I - \Pi_S) H \widetilde{f}(x)\|_2^2. \tag{H.9}$$

Choose $\widehat{S}$ by minimizing the average of ℓ_S over u independent inputs from Q. For every subsequent input, use the first coefficient of the least-squares fit on $\Phi_{\widehat{S}}$, clipped to $[0,1]$. A positive κ_H makes this target coefficient unique even if the nuisance coefficients are not.

Proposition H.2 (Recovery after common-support selection)**.** *For an input distribution $\mathcal{D}$, write $\mathcal{E}_H(\mathcal{D}) = \|H(\widetilde{f} - \mu)\|^2_{L^2(\mathcal{D})}$. Suppose $\kappa_H > 0$, $\widetilde{f}(x) \in [0,1]^L$, and $dP_X/dQ \le \rho$. Let $M = \binom{A}{s}$, $M_H = L\|H\|^2_{\text{op}}$, and*

$$t_u = M_H \sqrt{\frac{\log(2M/\alpha)}{2u}}.$$

Conditional on the fitted functions, with probability at least $1 - \alpha$ over the public inputs,

$$\|\widehat{p} - p\|^2_{L^2(P_X)} \le 2\kappa_H^{-2}\{\mathcal{E}_H(P_X) + \rho\, \mathcal{E}_H(Q) + 2\rho t_u\}. \tag{H.10}$$

Proof. Each residual lies in $[0, M_H]$. Hoeffding's inequality and a union bound control all empirical-to-population residual differences by t_u. The true support, enlarged to size s if needed, has residual at most $\|H(\widetilde{f} - \mu)\|_2^2$. Empirical minimization therefore gives $\mathbb{E}_Q \ell_{\widehat{S}} \le \mathcal{E}_H(Q) + 2t_u$. The density-ratio bound transfers this residual to P_X. In the weighted Euclidean norm, the distance from the fitted vector to $\mu(x)$ is at most $\sqrt{\ell_{\widehat{S}}(x)} + \|H(\widetilde{f}(x) - \mu(x))\|_2$. Apply target separation, the squared triangle inequality, and clipping to obtain (H.10). □

The conditional-prediction experiments use $H = (WW^\top)^{-1/2}$; all their model collections have linearly independent mixture rows. For numerical evaluation, probabilities are clipped to $[\epsilon, 1-\epsilon]$ with $\epsilon = 10^{-3}$. If p lies in this interval the bounds are unchanged; otherwise the squared-error bound becomes at most twice the stated bound plus $2\epsilon^2$ by the triangle inequality.

H.3 Choosing the amount of correction

Let $k \ge 1$, write $\mathcal{L}(f) = \|f - p\|^2_{L^2(P_X)}$, and condition on predictors fitted independently of the calibration sample. For two predictors $f_0, f_1 \in [0,1]$, continuous shrinkage uses $f_\lambda = f_0 + \lambda(f_1 - f_0)$ and minimizes empirical Brier loss on k independent unbiased target labels. With $\Delta_f = f_1 - f_0$,

$$\widehat{\lambda} = \operatorname{clip}_{[0,1]} \frac{\sum_{u=1}^k \Delta_f(X_u)(Y_u - f_0(X_u))}{\sum_{u=1}^k \Delta_f(X_u)^2}, \tag{H.11}$$

using zero when the denominator vanishes.

Proposition H.3 (Risk of empirical shrinkage)**.** *The rule* (H.11) *satisfies*

$$\mathbb{E}\mathcal{L}(f_{\widehat{\lambda}}) \le \min_{0\le\lambda\le 1} \mathcal{L}(f_\lambda) + \frac{5}{2\sqrt{k}}. \tag{H.12}$$

Proof. Put $\gamma_2 = \mathbb{E}\Delta_f(X)^2$ and $\gamma_1 = \mathbb{E}[\Delta_f(X)(f_0(X) - Y)]$. Apart from a common term, the population loss is $\lambda^2\gamma_2 + 2\lambda\gamma_1$. Empirical minimality and $0 \le \lambda, \widehat{\lambda} \le 1$ bound the excess loss by $|\widehat{\gamma}_2 - \gamma_2| + 2|\widehat{\gamma}_1 - \gamma_1|$. Since $\Delta_f^2 \in [0,1]$ and $\Delta_f(f_0 - Y) \in [-1,1]$, their sample means have expected absolute errors at most $1/(2\sqrt{k})$ and $1/\sqrt{k}$, respectively. □

The adaptive predictors in our experiments use ordinary averaging and source correction as f_0, f_1. Applying (H.7) to $f_{\widehat{\lambda}}$ then accounts for subsequent student approximation.

For a finite dictionary $f_1, \dots, f_{M'}$, classical exponential aggregation offers a different calibration rule (Dalalyan & Tsybakov, 2008). With prior weights $\varpi_j > 0$, $\sum_j \varpi_j = 1$, and a learning rate $0 < \beta \le 1/2$, set

$$w_{j,t} = \frac{\varpi_j \exp\{-\beta \sum_{u<t}(f_j(X_u) - Y_u)^2\}}{\sum_l \varpi_l \exp\{-\beta \sum_{u<t}(f_l(X_u) - Y_u)^2\}}, \quad g_t = \sum_j w_{j,t} f_j, \quad \bar{g} = k^{-1}\sum_{t=1}^{k} g_t. \tag{H.13}$$

For every j,

$$\mathbb{E}\mathcal{L}(\bar{g}) \le \mathcal{L}(f_j) + \frac{\log(1/\varpi_j)}{\beta k}. \tag{H.14}$$

Indeed, $z \mapsto \exp[-\beta(z-y)^2]$ is concave on $[0,1]$ for these β. Summing the resulting log-partition inequalities bounds cumulative regret by $\log(1/\varpi_j)/\beta$. Each g_t precedes its calibration observation, so its expected loss is its population loss. Convexity of squared loss transfers the average bound to $\bar{g}$, and subtracting the common irreducible Brier term proves (H.14).

H.4 Matched candidates for conditional predictions

The label-budget simulation and learned-model experiment use the single-task configuration in Table H.1, with $O_9 = \{M_1, \dots, M_7\}$ and $W = C/3$. Both candidates use three sources and have overlap vector $(1,1,1,2,1,1,1)$ with the existing models. All nine sources and their retained records are already used by O_9.

Table H.1: **Source records for the nine-source experiments.** Every model concerns the same target task. In the human-feedback experiment, sources 7 and 9 are affected by selection; G_9 includes both, while H_9 includes source 9.

Model	Sources	Model	Sources
M_1	$\{1,6,9\}$	M_5	$\{6,7,8\}$
M_2	$\{3,4,9\}$	M_6	$\{1,5,7\}$
M_3	$\{2,4,8\}$	M_7	$\{5,6,7\}$
M_4	$\{2,8,9\}$	G_9	$\{2,7,9\}$
		H_9	$\{5,8,9\}$

For O_9 and $O_9 \cup H_9$, the columns indexed by $\{3,4,7,9\}$ span $\mathbf{1}$. Enumerating sets of four columns of $W_{O_9 \cup G_9}$ gives $\operatorname{rank}([W_{:,T}, \mathbf{1}]) = \operatorname{rank}(W_{:,T}) + 1$ for every such T. Hence

$$\kappa_2(W_{O_9}) = \kappa_2(W_{O_9 \cup H_9}) = 0, \qquad \kappa_2(W_{O_9 \cup G_9}) > 0. \tag{H.15}$$

These candidates match in source counts and pairwise overlaps; the stronger unbiased joint-distribution matching is established separately for the configuration of Figure 1 in Appendix C.1.

To exhibit the ambiguity at the function level, take a nonconstant $p(x)$ bounded away from zero and one and a sufficiently small $d(x) > 0$. With u_a the ath coordinate vector, define two source-rate vectors

$$\begin{aligned} q(x) &= p(x)\mathbf{1} + 3d(x)(u_7 + u_9), \\ q'(x) &= [p(x) + d(x)]\mathbf{1} + 3d(x)(u_3 - u_4). \end{aligned} \tag{H.16}$$

Their target functions are p and $p + d$, respectively, and each has two biased sources. Set $v = (0, 0, -3, 3, 0, 0, 3, 0, 3)^\top$. Since $W_{O_9} v = \mathbf{1}$, $W_{H_9} v = 1$, and $W_{G_9} v = 2$, the population prediction mixtures satisfy

$$\mu_{O_9 \cup H_9}(x) = \mu'_{O_9 \cup H_9}(x), \qquad \mu_{G_9}(x) - \mu'_{G_9}(x) = d(x). \tag{H.17}$$

Thus G_9 distinguishes two target functions that all existing models and H_9 leave unresolved. These source functions are realizable by the input-dependent retention mechanism above. For separately trained delivery models, bounds (H.6) and (H.10) account for errors in fitting source predictors, approximating their mixtures, and selecting a common source support.