FROM BENCHMARK TO DEPLOYMENT: EFFICIENT EVALUATION OF LLMS FOR HEALTH UNDER TASK AND EVALUATOR SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) benchmarks evaluate models under a fixed task distribution and scoring protocol, both of which may differ from those relevant to deployment. We study how to reuse benchmark scores for deployment-specific model selection and performance estimation when scoring model responses on target tasks is resource-intensive. First, without target scores, we transport benchmark evaluations by reweighting tasks using semantic representations of prompts and rubrics, without requiring annotated shift attributes explicitly. On HealthBench, this approach can match or outperform reweighting based on known shift attributes in target-ranking recovery and mean-score estimation. Second, when a small number of target tasks are evaluated by the original *benchmark evaluator*, we combine their scores with benchmark scores to improve target-mean estimation and model-ranking recovery. Third, we address evaluator shift, a practical challenge when deployment evaluation relies on a human expert or a newer or more scalable LLM judge instead of the *benchmark evaluator*. Paired scores on a small target sample allow us to exploit the association between two evaluators' scores to improve target-mean estimation. We derive efficient influence functions for the target mean score in both the same- and different-evaluator settings. Building on these, we develop the Covariate-Adaptive Fusion Estimator (CAFE). Rather than separately estimating complex nuisance functions and plugging them into the efficient influence function, CAFE directly learns how to combine benchmark and target scores through variance minimization. Under suitable conditions, CAFE achieves semiparametric efficiency and supports asymptotically valid inference. Our experiments show that it improves target-mean estimation over existing transport and calibration approaches. Together, these results provide a principled route from static benchmark scores toward deployment-specific model selection and performance assessment under limited evaluation budgets.

1 INTRODUCTION

Expert-curated benchmarks such as HealthBench (Arora et al., 2025) provide a standardized basis for comparing large language models (LLMs) for health over a fixed collection of tasks, rubrics, and evaluation protocols. However, benchmark performance may not reflect performance in a deployment-relevant population: shifts in the task distribution can change average scores and even model rankings (Siska et al., 2024). Deployment evaluation may also use a different evaluator, whose judgments can differ systematically from those of the *benchmark evaluator* (Zheng et al., 2023). The resulting evaluation problem is how to reuse evidence from an existing benchmark when representative target cases are available but scoring them at scale is computationally intensive. This problem is especially relevant in high-stakes settings such as healthcare, where collecting target cases may be substantially easier than obtaining large-scale expert annotations. Recent discussions by the U.S. Food and Drug Administration similarly emphasize the limitations of static benchmarks for representing real-world conditions and consider evaluation frameworks that combine non-clinical benchmarking with clinical confirmation (U.S. Food and Drug Administration, 2026).

We first consider task-distribution shift with the evaluator held fixed, as illustrated in Figure 1(A), where target tasks are available but their scores are not. Reweighting can transport benchmark

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.

evaluations under covariate shift (Sugiyama et al., 2007), but the attributes driving the shift may not be explicitly annotated. Even when available, broad categories such as theme and language may capture only part of the task variation relevant to model performance. Prior work establishes representation-based weighting for covariate-shift correction and model evaluation (Stojanov et al., 2019; Chen et al., 2021). Building on these methods, we investigate whether the semantic information in prompts and evaluation rubrics can support benchmark transport without explicitly annotated target shift attributes. We construct representations from prompts alone or jointly from prompts and rubrics. These representations are used to estimate density ratios for reweighting benchmark tasks toward the target distribution. Prompts describe what a model is asked to do, whereas rubrics specify what constitutes a successful response, providing complementary information for deployment-specific evaluation. Our experiments on HealthBench show that prompt embeddings improve both target-mean estimation and model-ranking recovery, with further gains from incorporating rubric embeddings. Without explicit annotations of the attributes driving the target shift, embedding-based reweighting can match or outperform reweighting based on the known allocation attributes. These findings demonstrate how task content and evaluation criteria can supply transport-relevant information without first reducing them to explicitly annotated task categories.

A richer setting arises when a limited evaluation budget allows model responses on a subset of target tasks to be scored, as illustrated in Figure 1(B, C). These observed target scores provide direct information about target performance and can be combined with source benchmark scores. We refer to the evaluator used for benchmark scoring as the *benchmark evaluator* and the evaluator used for target scoring as the *target evaluator*. In deployment, the *target evaluator* may be a human expert or a different LLM judge, such as a newer or more scalable model. Although LLM judges offer a scalable alternative to human evaluation (Liu et al., 2023b; Zheng et al., 2023), scores can vary systematically across evaluators (Zheng et al., 2023). This issue is particularly relevant for rubric-based benchmarks such as HealthBench, where criteria may contain gradable terms such as "excessive" or "good," leaving room for different evaluators to score the same response differently. Our HealthBench evaluation illustrates this issue directly: GPT-4.1 and Gemini Flash-Lite assign different mean scores to the same model responses, with the magnitude of the difference varying across response models (Figure 1, inset).

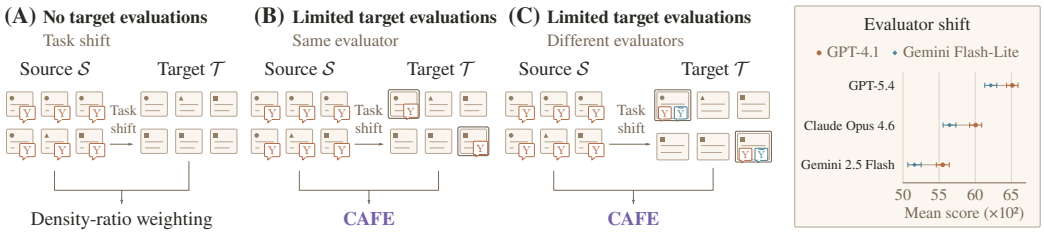

Figure 1: **Evaluation under task and/or evaluator shift.** **(A)** Without target evaluations, embedding-based reweighting transports benchmark scores to the target task distribution. **(B)** With limited target evaluations under the same evaluator, CAFE combines benchmark and target scores to improve estimation efficiency. **(C)** With limited target evaluations under a different evaluator, CAFE exploits residual association between paired benchmark- and target-evaluator scores to improve estimation efficiency. **Inset:** Mean HealthBench scores assigned by GPT-4.1 and Gemini Flash-Lite to responses from GPT-5.4, Claude Opus 4.6, and Gemini 2.5 Flash.

Prediction-powered methods provide a natural way to exploit auxiliary predictions when target scores are scarce (Angelopoulos et al., 2023a;b; Zrnic and Candès, 2024; Ji et al., 2025; Chen et al., 2026). Their efficiency gains are driven primarily by how informative those predictions are about the target scores. Our setting contains an additional source of information: source benchmark scores can contribute through structural relationships that are stable across the source and target populations. This motivates a data-fusion approach that combines transported benchmark-score information with limited target evaluations to estimate performance under the target distribution.

We develop the **Covariate-Adaptive Fusion Estimator (CAFE)** to combine benchmark and target scores for efficient estimation of target performance. We characterize semiparametric efficiency in both the same- and different-evaluator settings under conditional-mean stability of benchmark-evaluator scores across source and target populations. The resulting efficient influence function char-

acterizes when source benchmark scores provide information beyond the available target scores: in the different-evaluator setting, additional efficiency gains arise when there is residual association between benchmark- and target-evaluator scores after conditioning on the task representation. Implementing the efficient influence function directly would require estimating several complex nuisance functions including conditional variance and covariance functions. We avoid estimating these quantities separately by deriving a variance-minimization objective that directly learns how to combine benchmark and target scores. Across the HealthBench experiments, CAFE reduces target-mean estimation error in both the same- and different-evaluator settings. Model-ranking recovery also improves in the same-evaluator setting, while remaining broadly comparable to the target-only sample mean under evaluator shift. Our implementation and experimental code are available anonymously.[1]

2 Deployment Settings

Let X denote the task representation used for adjustment, constructed from the prompt and, when available, the rubric. We write $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$ for the source and target distributions, with corresponding expectations $\mathbb{E}_{\mathcal{S}}$ and $\mathbb{E}_{\mathcal{T}}$. We distinguish the availability of target evaluations from evaluator choice and consider three deployment settings.

In the setting with *no target evaluations*, Y denotes the *benchmark evaluator* score, observed only in the source population, and the target estimand is $\mu = \mathbb{E}_{\mathcal{T}}(Y)$. The observed data are $O_{\mathcal{S}} = (X, Y)$ and $O_{\mathcal{T}} = X$. To transport source scores to the target population, we assume conditional mean invariance, $\mathbb{E}_{\mathcal{S}}(Y \mid X) = \mathbb{E}_{\mathcal{T}}(Y \mid X)$, together with overlap, $p_{\mathcal{S}}(x) > 0$ whenever $p_{\mathcal{T}}(x) > 0$. Then $\mu = \mathbb{E}_{\mathcal{S}}\{q(X)Y\}$, where $q(x) = p_{\mathcal{T}}(x)/p_{\mathcal{S}}(x)$ is the target-to-source density ratio.

With *limited target evaluations under the same evaluator*, a subset of target tasks is also scored by the *benchmark evaluator* used in the source population. Let R indicate whether a target task is evaluated, with $R = 1$ when Y is observed, and assume random selection for evaluation, $\Pr_{\mathcal{T}}(R = 1 \mid X, Y) = \alpha$. The estimand remains $\mu = \mathbb{E}_{\mathcal{T}}(Y)$, and $O_{\mathcal{T}} = (X, R, RY)$.

Finally, with *limited target evaluations under a different evaluator*, Y denotes the *benchmark evaluator* score, while $\widetilde{Y}$ denotes the *target evaluator* score, and the estimand becomes $\mu = \mathbb{E}_{\mathcal{T}}(\widetilde{Y})$. Source tasks are scored by the *benchmark evaluator*, while a subset of target tasks receives paired benchmark and target evaluations. Thus, $R = 1$ indicates that both Y and $\widetilde{Y}$ are observed, with $\Pr_{\mathcal{T}}(R = 1 \mid X, Y, \widetilde{Y}) = \alpha$, and the observed data are $O_{\mathcal{S}} = (X, Y)$ and $O_{\mathcal{T}} = (X, R, RY, R\widetilde{Y})$. Throughout, O denotes an observed record, including the sample indicator $G \in \{\mathcal{S}, \mathcal{T}\}$. For each $g \in \{\mathcal{S}, \mathcal{T}\}$, let n_g denote the corresponding sample size, and define $N = n_{\mathcal{S}} + n_{\mathcal{T}}$ and $\pi_g = n_g/N$.

3 Related Work

LLM benchmark evaluation under distribution shift. Recent work increasingly recognizes that LLM benchmark performance depends on the distribution of tasks being evaluated, motivating benchmarks that better reflect real-world or deployment-facing use. Existing efforts construct evaluation distributions from live user interactions, naturally occurring conversations, or mixtures of benchmark tasks calibrated toward real-world query distributions (Chiang et al., 2024; Lin et al., 2025; Ni et al., 2024; Li et al., 2025b; Arora et al., 2025). Complementary studies examine how benchmark conclusions change when this distribution shifts, showing that model performance and comparisons can vary with task composition and with naturally occurring differences across domains, users, time, and geography (Kim et al., 2025; Seegmiller and Preum, 2026). These approaches primarily address benchmark–deployment mismatch by constructing, resampling, or directly evaluating on a more relevant target task population. A separate line of work studies evaluator mismatch, documenting systematic disagreement between LLM judges, humans, and alternative evaluators and motivating specialized judges, calibration to limited human evaluations, and statistical correction of noisy automatic scores (Zheng et al., 2023; Zeng et al., 2024; Tan et al., 2025; Kim et al., 2024; Kola et al., 2026; Chen et al., 2026). Related approaches combine imperfect annotations or predictions with probability-sampled gold labels, including design-based supervised learning and stratified prediction-powered inference (Egami et al., 2023; Fisch et al., 2024), while

[1] `https://anonymous.4open.science/r/Transport_HealthLLM-2838`

recent certification procedures account for uncertainty in estimated judge error rates (Feng et al., 2026). These methods primarily address evaluator mismatch while treating the evaluation-task distribution as fixed.

Data fusion and semiparametric efficiency. Transport methods leverage source outcomes under cross-population assumptions, often through density-ratio weighting or augmentation under covariate shift (Liu et al., 2023a; Chernozhukov et al., 2025). In high dimensions, unstable ratio estimation motivates low-dimensional, shift-relevant representations (Stojanov et al., 2019; Chen et al., 2021); Mandoline uses user-specified slicing functions. Our approach uses prompt and rubric text with density-ratio weighting to transport LLM benchmark scores without annotated shift attributes. Data-fusion theory studies how heterogeneous samples inform a common target under shared distributional components, weak alignment, or conditional-mean restrictions (Li and Luedtke, 2023; Qiu et al., 2024; Li et al., 2025a; Li, 2025; Graham et al., 2026). Closest in spirit, Li et al. (2023) improve efficiency by borrowing external outcomes under conditional-mean exchangeability. Our same-evaluator setting follows this principle, while the different-evaluator setting allows benchmark and target scores to come from different evaluators. Related work also studies efficiency gains from imperfect surrogates with limited primary outcomes (Kallus and Mao, 2025).

When target outcomes are observed only for a random subset, augmented inverse probability weighted (AIPW) and semi-supervised estimators combine outcome regression over all target covariates with residual correction from labelled observations (Robins et al., 1994; Zhang et al., 2019; Xu et al., 2025). Prediction-powered inference (PPI) follows a related principle, replacing or augmenting outcome regression with machine-learned predictions and using evaluated-target residuals for validity and efficiency (Angelopoulos et al., 2023a;b; Zrnic and Candès, 2024; Ji et al., 2025). Recent work further extends this framework through calibration and more general prediction-assisted estimators (van der Laan and van der Laan, 2026; Zou et al., 2026). The precision gain from predictions depends on how informative they are about target scores. In our setting, *benchmark evaluator* scores provide additional response-level information beyond prompt–rubric features, allowing source data to improve efficiency through residual association with the *target evaluator*.

4 Deployment under Task Shift

We test whether prompt- and rubric-based reweighting improves target-mean estimation and model-ranking recovery without target evaluations across three task-shift scenarios. Earlier work estimates target classifier performance without ground-truth annotations in the target domain using predictive confidence (Guillory et al., 2021; Garg et al., 2022) or sparse models of joint covariate–label shift (Chen et al., 2022).

Design and methods. We construct the three source–target shifts using 5,000 multilingual HealthBench examples spanning seven themes. HealthBench provides both prompts, which describe task content and context, and rubrics, which specify the criteria used to evaluate responses, allowing us to construct and compare prompt- and rubric-based representations. We evaluate five response models using their per-example HealthBench scores with GPT-4.1 as the fixed evaluator: GPT-5.4, Gemini 3.1 Pro, Gemini 3.1 Pro (Large Context), Claude Opus 4.6, and Gemini 2.5 Flash (Arora et al., 2025). Cases 1 and 2 increase the prevalence of four themes in the target population relative to the source: emergency referrals, health data tasks, global health, and context seeking. Case 1 induces a larger shift than Case 2. Case 3 jointly shifts theme and language composition, motivated by evidence that LLM performance in healthcare can vary substantially across languages and that English-centric evaluation may therefore underrepresent performance for non-English users (Jin et al., 2024). All methods are evaluated on the same 500 random splits per case, with target scores withheld from fitting and estimation. Appendix A.2 reports the allocation fractions and the corresponding source and target sample sizes.

As reference methods, *Baseline* uses unweighted source means, and *Oracle* is a Hájek estimator using known allocation weights by theme in Cases 1–2 and joint theme–language strata in Case 3. *Kernel mean matching (KMM)* (Huang et al., 2007) uses task embeddings to learn source weights by minimizing the discrepancy between the weighted source and target kernel mean embeddings. Neither model scores nor true target theme labels are used for fitting or tuning. For *KMM*, we compare prompt-only representations (Prompt) with concatenated prompt and rubric representa-

tions (Prompt+Rubric), using a fixed BGE-M3 encoder (Chen et al., 2024). The main analysis uses principal component analysis (PCA) to retain 50% of the variance separately within the prompt and rubric embedding blocks, yielding the PC50 representations. Implementation details appear in Appendix A.3. Appendices A.3–A.4 also describe additional representations and reweighting methods.

Metrics. We assess ranking using full rank recovery and Kendall's τ, and mean-score estimation using mean absolute error (MAE) and root mean squared error (RMSE) across the five model-specific target means per split (Figure 2; definitions in Appendix A.5). Appendix A.6 reports effective sample size (ESS), normalized by source sample size, to assess weight concentration.

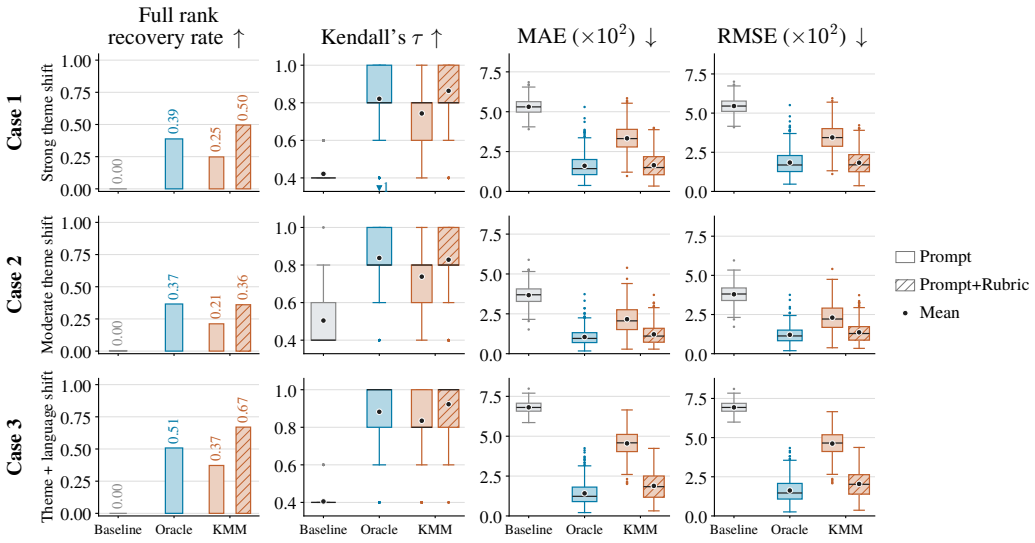

Figure 2: Comparison of reweighting methods under Cases 1–3 using PC50 representations. Boundary triangles and adjacent counts indicate observations outside the displayed range.

Prompt and rubric embeddings improve ranking and score estimation. Across all three settings, prompt-based *KMM* substantially improves both model-ranking recovery and mean-score estimation over the *Baseline*, although it remains less accurate than *Oracle*. Incorporating rubric embeddings further improves both outcomes, matching or surpassing *Oracle* without explicit annotations of the shift-driving attributes. One possible explanation is that *KMM* reduces covariate imbalance within allocation strata, where the *Oracle* weights are constant. This is analogous to the established result that inverse probability weighting with an estimated propensity score can be more efficient than weighting with the true propensity score (Hirano et al., 2003), illustrating why known allocation probabilities need not yield the most efficient estimates.

5 CAFE: Covariate-Adaptive Fusion Estimator

With limited target evaluations, source benchmark scores can improve estimation efficiency. We derive the efficient influence function (EIF) in both evaluator settings and use it to construct CAFE.

5.1 Semiparametric Efficiency Theory

In the same-evaluator setting, the observed target scores identify the target mean, while the source sample can provide additional information to improve efficiency. Borrowing this information requires a transportability condition linking the source and target populations. Existing semiparametric data-fusion and dataset-shift frameworks often rely on equality of conditional distributions across populations (Li and Luedtke, 2023; Qiu et al., 2024). Such distributional alignment may be unnecessarily strong in our setting. Since X is a low-dimensional task representation, it may not capture all factors governing the evaluation score, so the conditional score distributions may differ across populations even after conditioning on X. Motivated by mean-exchangeability approaches

that borrow external outcome information for efficiency gains (Li et al., 2023), we instead assume the following conditional mean restriction.

Assumption 1 (Conditional mean invariance)**.** $\mathbb{E}_{\mathcal{S}}[Y \mid X] = \mathbb{E}_{\mathcal{T}}[Y \mid X]$.

The different-evaluator setting retains this transportability condition for the benchmark-evaluator score Y, but allows the target score $\widetilde{Y}$ to be generated by a different evaluator. Thus, we do not require the two evaluators to have the same conditional mean. Instead, source observations of Y provide additional information about its shared conditional mean across populations, while paired $(Y, \widetilde{Y})$ observations in the evaluated target sample reveal how residual variation in the benchmark-evaluator score is related to residual variation in the target score. This residual association determines how much additional efficiency can be gained from the source sample.

We study semiparametric efficiency under both settings. The same-evaluator setting corresponds to $Y = \widetilde{Y}$, whereas the different-evaluator setting allows the two scores to differ. We derive the EIF of the target mean in both cases and use these characterizations to construct CAFE, which learns how much source information to borrow for efficient estimation.

Let $m(X) = \mathbb{E}_{\mathcal{S}}[Y \mid X] = \mathbb{E}_{\mathcal{T}}[Y \mid X]$ and let $\widetilde{m}(X) = \mathbb{E}_{\mathcal{T}}[\widetilde{Y} \mid X]$. Write $\sigma^2_{g,Y}(X) = \mathrm{Var}_g(Y \mid X)$ for $g \in \{\mathcal{S}, \mathcal{T}\}$, $\sigma^2_{\mathcal{T},\widetilde{Y}}(X) = \mathrm{Var}_{\mathcal{T}}(\widetilde{Y} \mid X)$, and $\gamma_{\mathcal{T}}(X) = \mathrm{Cov}_{\mathcal{T}}(Y, \widetilde{Y} \mid X)$. The following proposition characterizes the EIF of the target mean under the same- and different-evaluator settings.

Proposition 1. (i) Same-evaluator setting. *Assume Assumption 1 holds. The EIF of μ is*

$$\begin{aligned}\phi_1(O) = \frac{\mathbb{I}(G=\mathcal{T})}{\pi_{\mathcal{T}}}\{m(X)-\mu\} \\ + \frac{q(X)\{\mathbb{I}(G=\mathcal{S})\sigma^2_{\mathcal{T},Y}(X) + \mathbb{I}(G=\mathcal{T})R\sigma^2_{\mathcal{S},Y}(X)\}}{\pi_{\mathcal{S}}\sigma^2_{\mathcal{T},Y}(X) + \alpha\pi_{\mathcal{T}}q(X)\sigma^2_{\mathcal{S},Y}(X)}\{Y - m(X)\}.\end{aligned} \tag{1}$$

(ii) Different-evaluator setting. *Assume Assumption 1 holds. The EIF of μ is*

$$\begin{aligned}\phi_2(O) = \frac{\mathbb{I}(G=\mathcal{T})}{\pi_{\mathcal{T}}}\left[\widetilde{m}(X) - \mu + \frac{R}{\alpha}\{\widetilde{Y} - \widetilde{m}(X)\}\right] \\ + \left[\frac{\mathbb{I}(G=\mathcal{S})q(X)}{\pi_{\mathcal{S}}} - \frac{\mathbb{I}(G=\mathcal{T})R}{\alpha\pi_{\mathcal{T}}}\right]a^*(X)\{Y - m(X)\},\end{aligned} \tag{2}$$

where $a^*(x) = \pi_{\mathcal{S}}\gamma_{\mathcal{T}}(x)/\{\pi_{\mathcal{S}}\sigma^2_{\mathcal{T},Y}(x) + \alpha\pi_{\mathcal{T}}q(x)\sigma^2_{\mathcal{S},Y}(x)\}$.

Remark 1. In the same-evaluator setting, the first term of (1) reflects variation in the target covariate distribution. The second term combines residual information from the source and evaluated target samples, with weights determined by their relative sampling density and conditional variance. Thus, the EIF adaptively borrows more from the population that is locally more informative: when the source conditional variance is larger, the source residual receives less weight, and vice versa.

In the different-evaluator setting, equation (2) can be viewed as the target-only AIPW influence function under random selection of target tasks for evaluation (Robins et al., 1994), augmented by a correction based on the benchmark-evaluator residual $Y - m(X)$. The evaluated target sample reveals how this residual is associated with the target residual $\widetilde{Y} - \widetilde{m}(X)$, while the larger source sample provides additional observations of the benchmark-evaluator residual that can be transported to the target population. Accordingly, relative to the target-only AIPW estimator, which ignores the source sample, the efficiency gain is $\mathbb{E}_{\mathcal{T}}\{a^*(X)\gamma_{\mathcal{T}}(X)\}/\{\alpha\pi_{\mathcal{T}}\}$. The gain is strictly positive whenever $\gamma_{\mathcal{T}}(X) \neq 0$ on a set of positive target probability. In other words, after accounting for the task representation, the two evaluators' scores remain correlated for a non-negligible subset of target tasks. When $\gamma_{\mathcal{T}}(X) = 0$, $\phi_2(O)$ reduces to the target-only AIPW influence function.

5.2 Covariate-Adaptive Fusion Estimator

The EIFs $\phi_1(O)$ and $\phi_2(O)$ provide the efficiency benchmarks for estimating the target mean μ. Since $\phi_1(O)$ is recovered from $\phi_2(O)$ by setting $\widetilde{Y} = Y$ and $\widetilde{m} = m$, we present the estimator for the general different-evaluator setting; the same construction applies when the evaluators coincide.

A direct EIF-based estimator would replace the nuisance functions in $\phi_2(O)$ by estimates and solve $\mathbb{P}_N \phi_2(O; \mu, \widehat{\Theta}) = 0$, where $\Theta = \{\widetilde{m}, m, a^*, q, \alpha\}$ and $\mathbb{P}_N$ denotes the empirical average. The main challenge is estimating a^*. The expression for a^* in Proposition 1 suggests separate estimation of the conditional covariance $\gamma_{\mathcal{T}}$, two conditional variances, and the density ratio q. With limited paired target evaluations, the conditional second moments can be difficult to estimate accurately, and directly forming their ratio can be unstable (Young and Shah, 2024; Dong et al., 2026).

CAFE avoids this direct plug-in step by exploiting a variational characterization of a^*. Let $\phi_2(O; a)$ denote (2) with a^* replaced by a measurable coefficient function a. Under the conditional-mean invariance assumption, $\phi_2(O; a)$ forms a class of influence functions of μ, with the EIF obtained at $a = a^*$. Since the contribution from $\widetilde{m}(X) - \mu$ does not depend on a and is orthogonal to the residual contributions, minimizing $\mathrm{Var}\{\phi_2(O; a)\}$ over a is equivalent to minimizing

$$Q(a) = \frac{1}{2}\mathbb{E}_{\mathcal{T}}\left[\{\widetilde{Y} - \widetilde{m}(X) - a(X)(Y - m(X))\}^2\right] + \frac{\alpha\pi_{\mathcal{T}}}{2\pi_{\mathcal{S}}}\mathbb{E}_{\mathcal{S}}\left[q(X)^2 a(X)^2 \{Y - m(X)\}^2\right],$$

whose population minimizer is exactly a^*. This characterization suggests estimating a^* by minimizing the empirical counterpart of $Q(a)$. We optimize over a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, using an RKHS-norm penalty to control complexity of $a(X)$ and stabilize estimation when paired target evaluations are limited. Although the evaluation probability α is fixed by design, we estimate it by the empirical evaluated fraction $\widehat{\alpha}$ in the sample used to fit the objective. Given estimates $\widehat{m}$, $\widehat{\widetilde{m}}$, $\widehat{\alpha}$, and $\widehat{q}$, define

$$\widehat{Q}(a) = \frac{1}{2}\mathbb{P}_{\mathcal{T}}\left[\frac{R}{\widehat{\alpha}}\{\widetilde{Y} - \widehat{\widetilde{m}}(X) - a(X)(Y - \widehat{m}(X))\}^2\right] + \frac{\widehat{\alpha}\pi_{\mathcal{T}}}{2\pi_{\mathcal{S}}}\mathbb{P}_{\mathcal{S}}\left[\widehat{q}(X)^2 a(X)^2 \{Y - \widehat{m}(X)\}^2\right]. \tag{3}$$

We estimate a^* by $\widehat{a} \in \arg\min_{a \in \mathcal{H}} \left\{\widehat{Q}(a) + \frac{\lambda_a}{2}\|a\|_{\mathcal{H}}^2\right\}$, where λ_a is a regularization parameter.

Because the density ratio q depends only on the observed covariates and population membership, we estimate it using all source and target covariates and keep $\widehat{q}$ fixed across folds. To accommodate flexible estimation of the other nuisance functions, we use cross-fitting (Chernozhukov et al., 2018). Specifically, we partition the source and target samples into L folds; for each fold, the regressions, evaluation probability, and a are estimated using the remaining observations. These fitted quantities are then used to compute the estimator on the held-out fold, and the fold-specific estimates are averaged to obtain the final estimator. Algorithm 1 summarizes the procedure, with detailed construction and variance estimation provided in Appendix B. Theorem 1 establishes the asymptotic properties of the resulting estimator.

Algorithm 1 CAFE

Input: Source data $\mathcal{D}_{\mathcal{S}}$, target data $\mathcal{D}_{\mathcal{T}}$, and number of folds $L \geq 2$.
Output: Target-mean estimate $\widehat{\mu}$.
1: Estimate the density ratio $\widehat{q}$ using all source and target covariates.
2: Partition the source and target samples into L folds.
3: **for** $\ell = 1, \ldots, L$ **do**
4: Using the training folds, estimate $\widehat{\alpha}_\ell$, $\widehat{m}_\ell$, and $\widehat{\widetilde{m}}_\ell$.
5: Obtain $\widehat{a}_\ell$ by regularized minimization of (3), and define $\widehat{w}_\ell = \widehat{a}_\ell \widehat{q}$.
6: Evaluate the source and target score contributions on the corresponding held-out folds.
7: **end for**
8: Aggregate the held-out score contributions across folds and populations to obtain $\widehat{\mu}$.
9: **return** $\widehat{\mu}$.

Theorem 1 (Informal)**.** *Assume Assumption 1 holds. Under some regularity and rate-robustness conditions,* $\sqrt{N}(\widehat{\mu} - \mu) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \phi_2(O_i) + o_P(1) \xrightarrow{d} \mathcal{N}(0, V)$*, where* $V = \lim_N \mathrm{Var}\{\phi_2(O)\}$.

The proof is provided in Appendix D.3. For the same-evaluator setting, we provide the corresponding algorithm in Appendix C.

6 Experiments

In this section, we evaluate CAFE in both the same-evaluator and different-evaluator settings with limited target evaluations. We first compare CAFE with target-only and data-fusion baselines using semi-synthetic data. We then compare methods across the three HealthBench cases introduced in Section 4.

6.1 Evaluation on Semi-synthetic data

In this section, we compare CAFE with target-only and data-fusion baselines. We generate semi-synthetic data to retain the task-distribution structure of HealthBench while ensuring that the generated scores satisfy Assumption 1. In the different-evaluator setting, we vary the marginal Pearson correlation between the auxiliary score Y and the target score $\widetilde{Y}$, denoted by $\rho_{Y,\widetilde{Y}}$, over $\{0.4, 0.6, 0.8\}$. The detailed data-generating process is described in Appendix E.1.1.

Baselines. We compare CAFE with several baselines. In both evaluation settings, we include the *target-only sample mean*, which averages the observed scores from evaluated target tasks; *target-only AIPW*, which combines an outcome regression fitted on evaluated target data with a target residual correction; and efficient prediction-powered inference (*PPI++*), which uses source-trained score predictions and calibrates their contribution using target scores. In the same-evaluator setting, we additionally include *pooled-label AIPW*, which fits the outcome regression using both source and evaluated target data, and the *pooled-label doubly robust (DR) estimator*, which also incorporates source residuals while adjusting for source–target covariate shift. In the different-evaluator setting, we additionally evaluate recalibrated prediction-powered inference (*RePPI*), which learns a target-specific recalibration of the source-trained predictions. Within each setting, all methods use the same target evaluation assignments, and all outcome regressions are fitted using ridge regression. CAFE, target-only AIPW, pooled-label AIPW, and pooled-label DR use five-fold cross-fitting. Implementation details are provided in Appendices E.2.2 and E.3.2.

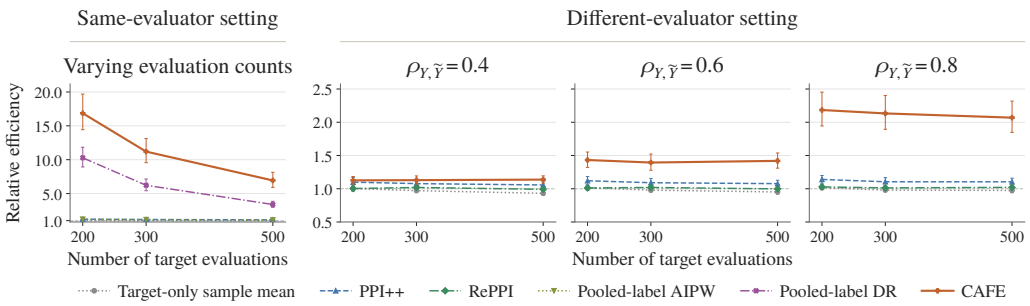

Figure 3: Relative efficiency, defined as the estimation-error variance of *target-only AIPW* divided by that of each method; values above one indicate greater efficiency. Error bars show 95% Monte Carlo intervals based on 500 splits.

Results. Figure 3 shows that, in the same-evaluator setting, CAFE achieves the largest efficiency gains over *target-only AIPW*, especially with fewer target scores. *Pooled-label DR* also improves efficiency, but to a lesser extent, while the remaining methods perform similarly to *target-only AIPW*. In the different-evaluator setting, CAFE shows modest gains over *target-only AIPW* when $\rho_{Y,\widetilde{Y}} = 0.4$, with substantial gains at larger $\rho_{Y,\widetilde{Y}}$. For context, the GPT-4.1 and Gemini Flash-Lite scores for Gemini 3.1 Pro responses in HealthBench have $\rho_{Y,\widetilde{Y}} = 0.793$, making the $\rho_{Y,\widetilde{Y}} = 0.8$ setting particularly relevant to our application. Competing methods remain close to *target-only AIPW*. Bias, MAE, and RMSE are reported in Tables 5–8 in Appendix E.1.2.

6.2 Evaluation on HealthBench

Experimental design. For each HealthBench setting considered in Section 4, we evaluate all five response models using 200, 300, or 500 evaluated target tasks, with scores unavailable for the remaining target tasks. In the same-evaluator setting, GPT-4.1 serves as the evaluator. In the different-evaluator setting, GPT-4.1 serves as the *benchmark evaluator* and Gemini Flash-Lite as the *target evaluator*. Throughout, we construct X using 58 prompt components and 22 rubric components, which explain 50% of the variance of prompt and rubric embeddings, respectively. Implementation details for CAFE are provided in Appendix E.2.1 for the same-evaluator setting and E.3.1 for the different-evaluator setting.

Table 1: Case 2 results in the same- and different-evaluator settings. Columns give target-evaluation counts; parentheses in method names identify samples supplying observed scores, with $\mathcal{L} = \{j, R_j = 1\}$ indexing evaluated target tasks. Green/blue denote ranks 1/2 within each setting.

Method	Full rank recovery ↑			Kendall's τ ↑			MAE ($\times 10^2$) ↓			RMSE ($\times 10^2$) ↓		
	200	300	500	200	300	500	200	300	500	200	300	500
(A) Same-evaluator setting												
Target-only sample mean ($\mathcal{L}$)	0.28	0.33	0.40	0.78	0.81	0.84	1.66	1.33	0.99	1.89	1.51	1.13
Source reweighting ($\mathcal{S}$)	0.36	0.36	0.36	0.83	0.83	0.83	1.21	1.21	1.21	1.36	1.36	1.36
PPI++ ($\mathcal{S}, \mathcal{L}$)	0.27	0.32	0.39	0.77	0.80	0.84	1.58	1.24	0.94	1.82	1.43	1.08
Target-only AIPW ($\mathcal{L}$)	0.29	0.32	0.37	0.78	0.80	0.83	1.64	1.30	0.98	1.88	1.49	1.12
Pooled-label AIPW ($\mathcal{S}, \mathcal{L}$)	0.29	0.32	0.37	0.78	0.81	0.83	1.57	1.24	0.94	1.81	1.43	1.08
Pooled-label DR ($\mathcal{S}, \mathcal{L}$)	0.42	0.39	0.42	0.84	0.84	0.86	1.01	0.93	0.82	1.15	1.05	0.93
CAFE ($\mathcal{S}, \mathcal{L}$)	0.43	0.42	0.45	0.86	0.86	0.87	0.95	0.86	0.75	1.09	0.99	0.85
(B) Different-evaluator setting												
Target-only sample mean ($\mathcal{L}$)	0.48	0.54	0.56	0.87	0.89	0.91	1.66	1.38	1.04	1.90	1.56	1.18
Source reweighting ($\mathcal{S}$)	0.31	0.31	0.31	0.81	0.81	0.81	3.71	3.71	3.71	3.91	3.91	3.91
PPI++ ($\mathcal{S}, \mathcal{L}$)	0.48	0.51	0.56	0.87	0.89	0.91	1.60	1.31	0.98	1.84	1.49	1.12
RePPI ($\mathcal{S}, \mathcal{L}$)	0.44	0.53	0.54	0.87	0.89	0.91	1.65	1.36	1.00	1.90	1.55	1.15
Target-only AIPW ($\mathcal{L}$)	0.46	0.52	0.55	0.87	0.89	0.91	1.67	1.35	1.01	1.91	1.54	1.16
CAFE ($\mathcal{S}, \mathcal{L}$)	0.46	0.53	0.52	0.88	0.90	0.90	1.36	1.13	0.91	1.58	1.31	1.05

Results in the same-evaluator setting. With only 200 target scores, CAFE outperforms the target-only sample mean using 500 target evaluations on all four metrics in Case 2 (Table 1). Compared with the target-only sample mean using the same 200 target evaluations, it increases full-rank recovery from 28% to 43% while reducing both MAE and RMSE by more than 40%. Thus, reusing benchmark scores improves both model selection and score estimation, even beyond what substantially more target evaluations achieve on their own. This advantage persists as the number of target scores increases: in Cases 1 and 2, CAFE matches or outperforms every baseline on all four metrics, including source-only reweighting (Tables 1 and 9). Under the joint theme and language shift in Case 3, CAFE achieves the lowest RMSE with 200, 300, and 500 target evaluations and the strongest ranking performance with 300 and 500 target evaluations (Table 10), showing that its benefits extend to more heterogeneous target populations.

Results in the different-evaluator setting. In Case 2, CAFE reduces MAE and RMSE relative to the target-only sample mean by approximately 18% and 17% using 200 target scores. It achieves the lowest MAE and RMSE across all three target-evaluation counts in Cases 1 and 2, and with 200 and 300 target evaluations in Case 3 (Tables 1, 11, and 12). Full-rank recovery varies across cases: it generally falls below that of the target-only sample mean in Cases 1 and 2, but matches or exceeds it in Case 3, reaching 62% versus 59% with 500 target evaluations.

7 Discussion

Our results show how benchmark information can support deployment-specific evaluation under task and evaluator shift. Without target evaluations or explicit annotations of shift-driving attributes, prompt- and rubric-based reweighting improves target-mean estimation and model-ranking recovery. With limited target evaluations, CAFE exploits residual association between benchmark- and target-evaluator scores beyond task representations to improve estimation efficiency. Thus, benchmark scores can remain informative even when task representations explain only a limited fraction of individual-score variation. Our HealthBench experiments support these findings in both same- and different-evaluator settings, highlighting the complementary roles of task representations, benchmark scores, and limited target evaluations.

These gains do not imply robustness to arbitrary unmeasured shifts: transport still requires adequate overlap and conditional-mean stability given the representation. Future work could extend the framework to benchmarks without explicit rubrics by learning evaluation-relevant structure from prompts, metadata, or generated criteria, and could study adaptive selection of target tasks for evaluation.

AI USE STATEMENT

Generative artificial intelligence tools assisted with mathematical exposition, manuscript organization, language editing, code organization, and summaries of experimental outputs. The authors are responsible for the mathematical claims, empirical results, citations, and final text.

REFERENCES

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

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b. https://arxiv.org/abs/2311.01453.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025. https://arxiv.org/abs/2505.08775.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(75):2137–2155, 2009. https://www.jmlr.org/papers/v10/bickel09a.html.

Mayee Chen, Karan Goel, Nimit S. Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Ré. Mandoline: Model evaluation under distribution shift. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1617–1629, 2021. https://proceedings.mlr.press/v139/chen21i.html.

Lingjiao Chen, Matei Zaharia, and James Zou. Estimating and explaining model performance when both covariates and labels shift. *Advances in Neural Information Processing Systems*, 35:11467–11479, 2022. https://doi.org/10.52202/068431-0833.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024. https://doi.org/10.18653/v1/2024.findings-acl.137.

Yiqun T. Chen, Sizhu Lu, Sijia Li, Moran Guo, and Shengyi Li. Efficient inference for noisy LLM-as-a-judge evaluation. In *International Conference on Machine Learning (ICML)*, 2026. https://icml.cc/virtual/2026/poster/61131.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. https://doi.org/10.1111/ectj.12097.

Victor Chernozhukov, Michael Newey, Whitney K. Newey, Rahul Singh, and Vasilis Syrgkanis. Automatic debiased machine learning for covariate shifts. *arXiv preprint arXiv:2307.04527*, 2025. https://arxiv.org/abs/2307.04527.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024. https://proceedings.mlr.press/v235/chiang24b.html.

Antoine de Mathelin, François Deheeger, Guillaume Richard, Mathilde Mougeot, and Nicolas Vayatis. ADAPT: Awesome Domain Adaptation Python Toolbox. *arXiv preprint arXiv:2107.03049*, 2021. https://arxiv.org/abs/2107.03049v1.

Mei Dong, Lin Liu, Dingke Tang, Geoffrey Liu, Wei Xu, and Linbo Wang. Average treatment effect estimation with non-binary instrumental variables. *arXiv preprint arXiv:2510.14368v2*, 2026. https://arxiv.org/abs/2510.14368v2.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, 36:68589–68601, 2023. https://doi.org/10.52202/075280-3000.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In *The Fourteenth International Conference on Learning Representations*, 2026. https://openreview.net/forum?id=hEhxreaLdU.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. *Advances in Neural Information Processing Systems*, 37:111489–111514, 2024. https://doi.org/10.52202/079017-3541.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *The Tenth International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=o_HsiMPYh_x.

Ellen Graham, Marco Carone, and Andrea Rotnitzky. Towards a unified theory for semiparametric data fusion with individual-level data. *The Annals of Statistics*, 54(3):1398–1424, 2026. https://doi.org/10.1214/25-AOS2609.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026. https://arxiv.org/abs/2605.13986.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1134–1144, 2021. https://openaccess.thecvf.com/content/ICCV2021/html/Guillory_Predicting_With_Confidence_on_Unseen_Distributions_ICCV_2021_paper.html.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003. https://doi.org/10.1111/1468-0262.00442.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. https://doi.org/10.1038/s41586-024-08328-6.

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems*, 19:601–608, 2007. https://proceedings.neurips.cc/paper_files/paper/2006/hash/a2186aa7c086b46ad4e8bf81e2a3a19b-Abstract.html.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of AI. *arXiv preprint arXiv:2501.09731*, 2025. https://doi.org/10.48550/arXiv.2501.09731.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. Better to ask in English: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, pp. 2627–2638, 2024. https://doi.org/10.1145/3589334.3645643.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431, 2017. https://aclanthology.org/E17-2068/.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016. https://arxiv.org/abs/1612.03651.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):480–509, 2025. https://doi.org/10.1093/jrsssb/qkae099.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009. https://www.jmlr.org/papers/v10/kanamori09a.html.

Eunsu Kim, Haneul Yoo, Guijin Son, Hitesh Patel, Amit Agarwal, and Alice Oh. BenchHub: A unified benchmark suite for holistic and customizable LLM evaluation. *arXiv preprint arXiv:2506.00482*, 2025. https://arxiv.org/abs/2506.00482.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024. https://doi.org/10.18653/v1/2024.emnlp-main.248.

Sneha Kola, Pankaj Kumar Sharma, Soumyadeep Dey, Bamdev Mishra, and Mayur Datar. SAJA: A simple approach to judge alignment for LLM-as-a-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 646–664, 2026. https://doi.org/10.18653/v1/2026.acl-industry.45.

Harrison H. Li. Efficient estimation and data fusion under general semiparametric restrictions on outcome mean functions. *arXiv preprint arXiv:2406.06941*, 2025. https://doi.org/10.48550/arXiv.2406.06941.

Sijia Li and Alex Luedtke. Efficient estimation under data fusion. *Biometrika*, 110(4):1041–1054, 2023. https://doi.org/10.1093/biomet/asad007. Correction: *Biometrika*, 112(2):asaf005, 2025. https://doi.org/10.1093/biomet/asaf005.

Sijia Li, Peter B. Gilbert, Rui Duan, and Alex Luedtke. Data fusion using weakly aligned sources. *Journal of the American Statistical Association*, 120(552):2569–2579, 2025a. https://doi.org/10.1080/01621459.2025.2476780.

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

Xinyu Li, Wang Miao, Fang Lu, and Xiao-Hua Zhou. Improving efficiency of inference in clinical trials with external control data. *Biometrics*, 79(1):394–403, 2023. https://doi.org/10.1111/biom.13583.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/771155abaae744e08576f1f3b4b7ac0d-Abstract-Conference.html.

Molei Liu, Yi Zhang, Katherine P. Liao, and Tianxi Cai. Augmented transfer regression learning with semi-non-parametric nuisance models. *Journal of Machine Learning Research*, 24(293):1–50, 2023a. https://www.jmlr.org/papers/v24/22-0700.html.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023b. https://doi.org/10.18653/v1/2023.emnlp-main.153.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. MixEval: Deriving wisdom of the crowd from LLM benchmark mixtures. *Advances in Neural Information Processing Systems*, 37:98180–98212, 2024. https://doi.org/10.52202/079017-3116.

Hongxiang Qiu, Eric Tchetgen Tchetgen, and Edgar Dobriban. Efficient and multiply robust risk estimation under general forms of dataset shift. *The Annals of Statistics*, 52(4):1796–1824, 2024. https://doi.org/10.1214/24-AOS2422.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. https://doi.org/10.1080/01621459.1994.10476818.

Parker Seegmiller and Sarah Masud Preum. Measuring distribution shift in user prompts and its effects on LLM performance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32685–32704, 2026. https://doi.org/10.18653/v1/2026.acl-long.1508.

Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bono. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10406–10421, 2024. https://doi.org/10.18653/v1/2024.acl-long.560.

Petar Stojanov, Mingming Gong, Jaime Carbonell, and Kun Zhang. Low-dimensional density ratio estimation for covariate shift correction. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3449–3458, 2019. https://proceedings.mlr.press/v89/stojanov19a.html.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007. https://www.jmlr.org/papers/v8/sugiyama07a.html.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008. https://doi.org/10.1007/s10463-008-0197-x.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations*, 2025. https://openreview.net/forum?id=G0dksFayVq.

U.S. Food and Drug Administration. Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback. Center for Devices and Radiological Health, U.S. Food and Drug Administration, 2026. https://www.fda.gov/media/194242/download.

Lars van der Laan and Mark van der Laan. Calibeating prediction-powered inference. *arXiv preprint arXiv:2604.21260*, 2026. https://doi.org/10.48550/arXiv.2604.21260.

Zichun Xu, Daniela Witten, and Ali Shojaie. A unified framework for semiparametrically efficient semi-supervised learning. *arXiv preprint arXiv:2502.17741*, 2025. https://doi.org/10.48550/arXiv.2502.17741.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013. https://doi.org/10.1162/NECO_a_00442.

Elliot H. Young and Rajen D. Shah. ROSE random forests for robust semiparametric efficient estimation. *arXiv preprint arXiv:2410.03471*, 2024. https://arxiv.org/abs/2410.03471.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/afc8b034823271816d14f7c1aefe1dff-Abstract-Conference.html.

Anru Zhang, Lawrence D. Brown, and T. Tony Cai. Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566, 2019. https://doi.org/10.1214/18-AOS1756.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Runjia Zou, Daniela Witten, and Brian Williamson. Generalized prediction-powered inference, with application to binary classifier evaluation. *arXiv preprint arXiv:2602.10332*, 2026. https://doi.org/10.48550/arXiv.2602.10332.

Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024. https://doi.org/10.1073/pnas.2322083121.

A TASK-SHIFT DETAILS

The task-shift experiments examine whether model evaluations obtained from a source benchmark can be adapted to a target population with a different mix of health tasks. We hold the model responses and the evaluator fixed and consider three settings involving changes in theme composition and language. We investigate how the choice of representation and reweighting method affects the recovery of target mean scores and model rankings, with particular attention to what rubric information adds beyond prompts. These two evaluation goals are considered separately because accurate model rankings need not imply accurate estimates of absolute performance.

A.1 DATA AND MODEL SCORES

We use all 5,000 multilingual examples from HealthBench (Arora et al., 2025). Each example provides a conversation prompt and an evaluation rubric specifying the criteria used to assess a model response. Each example belongs to one of seven task themes: emergency referrals, health data tasks, global health, context seeking, expertise-tailored communication, responding under uncertainty, and response depth. We use saved responses from five models: GPT-5.4, Gemini 3.1 Pro, Gemini 3.1 Pro (Large Context), Claude Opus 4.6, and Gemini 2.5 Flash, and these responses are judged and scored by GPT-4.1 using the corresponding evaluation rubrics. For response model k, let Y_{ik} denote the score assigned by the GPT-4.1 judge to its response to example i. The responses and judge scores remain fixed across the source–target splits described below.

GPT-5.4 has 4,999 observed scores, Claude Opus 4.6 has 4,991, and each remaining model has 5,000. All examples enter split construction in Appendix A.2 and representation fitting in Appendix A.3 regardless of score availability. Scores are not imputed, and mean calculations use the observed scores for the respective model, as specified in Appendix A.5.

A.2 SOURCE–TARGET SPLIT DESIGN

We construct three cases by splitting the same 5,000 multilingual HealthBench examples into source and target data. Cases 1 and 2 vary the relative frequencies of these themes, while Case 3 additionally changes the language composition within themes. For each case, we generate 500 splits. Target data is sampled uniformly without replacement within the specified strata, and the remaining examples form the source data. Within each split, source cases are indexed by $i = 1, \ldots, n_{\mathcal{S}}$ and target cases by $j = 1, \ldots, n_{\mathcal{T}}$.

Cases 1 and 2: theme shift. Case 1 creates a more pronounced difference in theme proportions between the source and target data, whereas Case 2 uses less extreme allocations, resulting in more similar theme proportions. Let n_h denote the number of examples in theme h and a_h its specified target allocation fraction. Table 2 reports these fractions and the resulting source and target sample sizes. For Cases 1 and 2, the target count in theme h is $n_{\mathcal{T},h} = \lfloor n_h a_h + 0.5 \rfloor$, and the remaining $n_{\mathcal{S},h} = n_h - n_{\mathcal{T},h}$ examples form the source set.

Table 2: Target allocation fractions a_h within each theme and total source and target sample sizes for Cases 1 and 2. Sample sizes are fixed across splits.

Theme	Case 1	Case 2
Emergency referrals	95%	90%
Health data tasks	95%	90%
Global health	95%	90%
Context seeking	95%	90%
Expertise-tailored communication	30%	50%
Responding under uncertainty	30%	50%
Response depth	30%	50%
Source sample size ($n_{\mathcal{S}}$)	1,778	1,439
Target sample size ($n_{\mathcal{T}}$)	3,222	3,561

Case 3: joint theme and language shift. Case 3 extends the theme-based design by allowing the target allocation fraction to differ between English and non-English examples within each theme. This introduces changes in language composition alongside changes in theme proportions. We use a pretrained fastText language identification model (Joulin et al., 2017; 2016) to identify the primary language of each prompt and group examples as English or non-English.

Here, h indexes a joint theme–language stratum, with n_h and a_h defined analogously. For strata with $a_h > 0$, we retain at least one example in both source and target data. The number of examples assigned to the target data within stratum h is

$$n_{\mathcal{T},h} = \begin{cases} 0, & a_h = 0, \\ \min\{n_h - 1, \max\{1, \lfloor n_h a_h + 0.5 \rfloor\}\}, & a_h > 0. \end{cases}$$

Table 3 reports the allocation fractions and resulting counts within each stratum. Each split contains 2,310 source and 2,690 target examples. The proportion of non-English examples is 7.19% in source data and 21.93% in target data. The two non-English examples in the Health data tasks theme are always assigned to source, as the target allocation fraction for this stratum is zero. The allocation fractions for Case 3 were chosen using existing model-score summaries, making this setting an exploratory rank-shift stress test.

Table 3: Target allocation fractions a_h and source and target sample sizes within each theme–language stratum for Case 3. Sample sizes are fixed across splits.

	English			Non-English		
Theme	a_h	Source	Target	a_h	Source	Target
Emergency referrals	90%	46	412	90%	2	22
Health data tasks	90%	47	428	0%	2	0
Global health	90%	75	674	90%	35	313
Context seeking	90%	54	484	90%	6	50
Expertise-tailored communication	5%	774	41	5%	99	5
Responding under uncertainty	5%	845	45	90%	18	163
Response depth	5%	303	16	90%	4	37
Total		2,144	2,100		166	590

A.3 Prompt and rubric representations

We compare prompt representations (P) with combined prompt and rubric representations ($P + R$) to examine whether information about the evaluation criteria improves reweighting. Here, $+$ denotes concatenation. The prompt provides information about the task presented to a response model, while the rubric specifies the criteria used to assess its response.

Both representations are constructed using the same fixed BGE-M3 text encoder (Chen et al., 2024),[2] with a maximum input length of 8,192 tokens. For each example, the prompt representation is obtained by concatenating its user messages and encoding the resulting text. For the rubric representation, each criterion is encoded separately and normalized to unit length. These criterion embeddings are averaged with equal weight, and the resulting vector is normalized to unit length. Both the prompt and rubric embeddings have 1,024 coordinates and unit Euclidean norm.

To examine how representation dimension affects reweighting, we consider the full embeddings and four reduced representations obtained by principal component analysis (PCA). PCA is fitted separately to the prompt and rubric embeddings using all 5,000 examples, with centering and full singular value decomposition. This step uses neither model scores nor theme labels. The two PCA transformations are fitted once and held fixed across all three cases and all repetitions.

PC30, PC50, PC80 and PC95 retain the smallest number of leading components explaining at least 30%, 50%, 80% and 95% of the variance within each embedding block, respectively. For $P + R$, the retained prompt and rubric components are concatenated after the separate PCA projections. Full

[2] We use `BAAI/bge-m3` at revision `5617a9f61b028005a4858fdac845db406aefb181`, loaded through the SentenceTransformers package.

uses the original embeddings without dimensionality reduction. Table 4 gives the resulting input dimensions.

Table 4: Dimensions of P and $P + R$ under each representation setting.

Representation	Prompt (P)	Prompt + rubric ($P + R$)
PC30	20	28
PC50	58	80
PC80	191	300
PC95	377	662
Full	1,024	2,048

A.4 Reweighting methods and fitting

Across the main and supplementary analyses, we evaluate eight reweighting methods and two reference estimators. The unweighted source mean serves as the baseline (*Baseline*), while a Hájek estimator using known allocation weights provides an oracle reference (*Oracle*).

Five reweighting methods use task embeddings directly. A source–target classifier estimates density ratios from domain-membership probabilities (*Domain classifier*) (Bickel et al., 2009). Kernel mean matching balances the weighted source and target kernel mean embeddings (*Embedding KMM*, labelled simply *KMM* in the main text and figures) (Huang et al., 2007). Unconstrained least-squares importance fitting estimates the density ratio by minimizing a squared-error criterion (*Embedding uLSIF*) (Kanamori et al., 2009). Relative unconstrained least-squares importance fitting instead estimates a relative density ratio (*Embedding RuLSIF*) (Yamada et al., 2013). The Kullback–Leibler importance estimation procedure estimates the density ratio by maximizing the average log density ratio over target examples, subject to normalization on the source sample (*Embedding KLIEP*) (Sugiyama et al., 2008).

Two methods construct weights using predicted themes. Hard predicted-theme reweighting assigns each target example its most probable theme and uses the resulting theme frequencies to construct source weights (*Hard predicted-theme*). Kernel mean matching applied to predicted-theme probability vectors balances these probabilities between source and target samples (*Predicted-theme KMM*). Finally, kernel mean matching applied to embeddings concatenated with predicted-theme probabilities balances both representations jointly (*Hybrid KMM*).

Let Z denote a selected representation from Appendix A.3. Within each split, write $Z_i^{\mathcal{S}}$, $i = 1, \ldots, n_{\mathcal{S}}$, and $Z_j^{\mathcal{T}}$, $j = 1, \ldots, n_{\mathcal{T}}$, for the representations of source and target data, respectively. For each split, learned weights are fitted and tuned using source and target representations, domain membership, and, where applicable, source theme labels. Model scores do not enter weight fitting or tuning. Target theme labels are reserved for split construction and the oracle reference, and the language-group indicator is not appended as a separate feature.

Reference estimators. The *unweighted source mean* assigns equal weights to all source examples. The *oracle* uses the known allocation design to construct weights within the strata defining the shift. In Cases 1 and 2, a source example in theme h receives raw weight $a_h/(1-a_h)$, using the configured fractions in Table 2. In Case 3, the oracle uses the following raw weight:

$$\widetilde{w}_i^{\text{oracle}} = \frac{n_{\mathcal{T},h_i^{\mathcal{S}}}}{n_{\mathcal{S},h_i^{\mathcal{S}}}}, \qquad i = 1, \ldots, n_{\mathcal{S}}, \tag{4}$$

where $h_i^{\mathcal{S}}$ identifies the joint theme–language stratum of source example i. The resulting oracle is a Hájek reference estimator based on the known allocation design.

Methods using embeddings. A Tabular Prior-data Fitted Network classifier (TabPFN-3) (Hollmann et al., 2025; Grinsztajn et al., 2026) is fitted to pooled source and target representations, with each example labelled by whether it belongs to the source or target set. Five folds stratified by domain produce out-of-fold probabilities. If $\widehat{e}_i^{\mathcal{S}}$ is the predicted probability that source example i

belongs to the target domain, its raw weight is

$$\widetilde{w}_i = \frac{n_{\mathcal{S}}}{n_{\mathcal{T}}} \frac{\widehat{e}_i^{\mathcal{S}}}{1 - \widehat{e}_i^{\mathcal{S}}}.$$

The other four methods operate directly on the selected representations through a Gaussian kernel. *KMM* balances the weighted source and target kernel means. *Embedding uLSIF* estimates a density ratio by least-squares fitting. *Embedding RuLSIF* estimates a relative density ratio with mixing parameter 0.1; its fitted weights are used directly, followed by the normalization described in Appendix A.5. *Embedding KLIEP* maximizes the average log density ratio over target examples, subject to the estimated ratio averaging to one over source examples.

Methods using predicted themes. These two methods use a TabPFN-3 classifier trained on source theme labels to predict themes from the selected representations. The classification task has seven theme categories in all three cases, including Case 3.

Hard predicted-theme (argmax) reweighting fits the classifier on the full source sample and assigns each target example its highest-probability theme. Let $\widehat{f}_{\mathcal{T}}^{\,\mathrm{hard}}(h)$ denote the fraction of target examples assigned to theme h, and let $\widehat{f}_{\mathcal{S}}(h)$ denote its observed source fraction. A source example with observed theme $H_i^{\mathcal{S}}$ receives raw weight

$$\widetilde{w}_i = \frac{\widehat{f}_{\mathcal{T}}^{\,\mathrm{hard}}(H_i^{\mathcal{S}})}{\widehat{f}_{\mathcal{S}}(H_i^{\mathcal{S}})}.$$

Predicted-theme KMM uses the classifier's seven-dimensional probability vectors as inputs to KMM. Five folds stratified by source theme produce out-of-fold probabilities for source examples. Each fitted classifier also predicts probabilities for every target example, and these target predictions are averaged across the five classifiers. KMM then balances the source and target probability vectors.

Hybrid KMM. Hybrid KMM combines the selected representations with predicted-theme probabilities. Source probabilities are obtained out of fold, and target probabilities are averaged across the five source-trained classifiers, using the same construction as predicted-theme KMM. The seven probability columns are appended to the corresponding representations on their original scale, giving inputs with $\dim(Z) + 7$ coordinates. KMM then balances the weighted source and target kernel means of these combined inputs.

Fitting and tuning. All TabPFN-3 classifiers use eight estimators, softmax temperature 0.9, and `float32` inference, with automatic adjustment of the estimator count disabled.

The kernel methods use the Gaussian kernel

$$K_\gamma(u, v) = \exp\{-\gamma \|u - v\|^2\}.$$

For each split and input, let σ be the median pairwise Euclidean distance among at most 2,000 pooled source and target observations, with subsampling determined by the split seed. The candidate values are

$$\gamma \in \left\{ \frac{0.25}{2\sigma^2}, \frac{1}{2\sigma^2}, \frac{4}{2\sigma^2} \right\}.$$

Distances are computed in the input space used by the method: embeddings, predicted-theme probabilities, or their concatenation.

uLSIF and RuLSIF use 200 kernel centers and jointly select the bandwidth and regularization parameter, with regularization candidates $\{0.001, 0.01, 0.1, 1, 10\}$. KLIEP uses 200 centers, five-fold bandwidth selection, and Frank–Wolfe optimization with at most 2,000 iterations.

KMM uses a five-fold partition of the source sample, retaining all target covariates in each fit. It selects the bandwidth with the highest mean training-fold unsupervised score, using ADAPT's `neg_j_score` scoring function (de Mathelin et al., 2021), and then refits on all source and target covariates. KMM uses all source examples, a raw-weight bound $B_{\max} = 1000$, the implementation's default weight-sum tolerance, and at most 100 solver iterations.

A.5 TARGET MEANS, RANKING, AND UNCERTAINTY SUMMARIES

For response model k, let $Y_{ik}^{\mathcal{S}}$ and $Y_{jk}^{\mathcal{T}}$ denote the saved GPT-4.1 scores on source example i and target example j, respectively. For each split, method, and input combination, the fitted raw weights are normalized to source mean one,

$$w_i = \frac{\widetilde{w}_i}{n_{\mathcal{S}}^{-1} \sum_{i'=1}^{n_{\mathcal{S}}} \widetilde{w}_{i'}}.$$

For model k, let $\mathcal{S}_k \subseteq \{1, \ldots, n_{\mathcal{S}}\}$ and $\mathcal{T}_k \subseteq \{1, \ldots, n_{\mathcal{T}}\}$ be the source and target index sets with observed scores. The estimated target mean and its observed target reference are

$$\widehat{\mu}_k^{\mathcal{T}} = \frac{\sum_{i \in \mathcal{S}_k} w_i Y_{ik}^{\mathcal{S}}}{\sum_{i \in \mathcal{S}_k} w_i}, \qquad \overline{Y}_k^{\mathcal{T}} = \frac{1}{|\mathcal{T}_k|} \sum_{j \in \mathcal{T}_k} Y_{jk}^{\mathcal{T}}. \tag{5}$$

Thus each model uses its observed examples, with the weighted denominator recomputed over $\mathcal{S}_k$ rather than retaining $n_{\mathcal{S}}$ or the full-source weight sum. Scores are not imputed. The same saved weight vector is applied to all five response models for a given split, method, and input combination. The baseline is equation (5) with $w_i = 1$. With complete scores, the Case 3 oracle reduces to the target-stratum-weighted source mean,

$$\widehat{\mu}_k^{\mathcal{T},\text{oracle}} = \sum_h \frac{n_{\mathcal{T},h}}{n_{\mathcal{T}}} \overline{Y}_{h,k}^{\mathcal{S}},$$

where

$$\overline{Y}_{h,k}^{\mathcal{S}} = \frac{1}{n_{\mathcal{S},h}} \sum_{\substack{i=1 \\ h_i^{\mathcal{S}} = h}}^{n_{\mathcal{S}}} Y_{ik}^{\mathcal{S}}.$$

With missing scores, it instead uses the model-specific numerator and denominator in equation (5). The evaluation reference is the mean of observed target GPT-4.1 scores for that response model.

For split b and $K_{\text{mod}} = 5$ models, we evaluate the methods in terms of the following MAE and RMSE

$$\text{Avg.MAE}_b = \frac{1}{K_{\text{mod}}} \sum_{k=1}^{K_{\text{mod}}} \left| \widehat{\mu}_k^{\mathcal{T},(b)} - \overline{Y}_k^{\mathcal{T},(b)} \right|,$$

$$\text{Avg.RMSE}_b = \left\{ \frac{1}{K_{\text{mod}}} \sum_{k=1}^{K_{\text{mod}}} \left(\widehat{\mu}_k^{\mathcal{T},(b)} - \overline{Y}_k^{\mathcal{T},(b)} \right)^2 \right\}^{1/2}.$$

Throughout the paper and appendix, MAE and RMSE are reported in percentage points, on a 100-point score scale. Ranking accuracy is assessed by Kendall's τ between the five estimated and observed target means and by exact full-rank recovery. Full-rank recovery requires equality of all five descending ranks, retaining exact ties through competition ranks.

Also, the effective source-sample fraction is displayed as a percentage and is computed from all source weights:

$$\frac{\text{ESS}}{n_{\mathcal{S}}} = \frac{\left(\sum_{i=1}^{n_{\mathcal{S}}} w_i\right)^2}{n_{\mathcal{S}} \sum_{i=1}^{n_{\mathcal{S}}} w_i^2}, \qquad n_{\mathcal{S}} = 1{,}778,\ 1{,}439,\ 2{,}310 \quad \text{for Cases 1, 2, and 3, respectively.}$$

A.6 COMPARISON ACROSS METHODS AND REPRESENTATIONS

Figures 4–18 compare eight reweighting methods and two reference estimators for Cases 1–3 using PC30, PC50, PC80, PC95, and full embeddings.

Across different cases and representations, the additional results support the main-text findings. KMM-based methods, particularly *embedding KMM* and *hybrid KMM*, are generally among the strongest reweighting methods for both ranking recovery and target-mean estimation. In contrast, the domain-classification, density-ratio-estimation, and predicted-theme alternatives show less consistent improvements across settings. Performance also tends to deteriorate with very high-dimensional

representations, while the lower-dimensional PCA representations provide a more stable trade-off between adjustment and weight variability. Across methods, incorporating rubric information generally improves score-estimation accuracy beyond prompt-only representations. The ESS results provide a complementary diagnostic of weight concentration. Higher-dimensional representations often lead to lower effective source sample sizes for some learned weighting methods, consistent with less stable weighting, although ESS alone does not fully explain estimation accuracy.

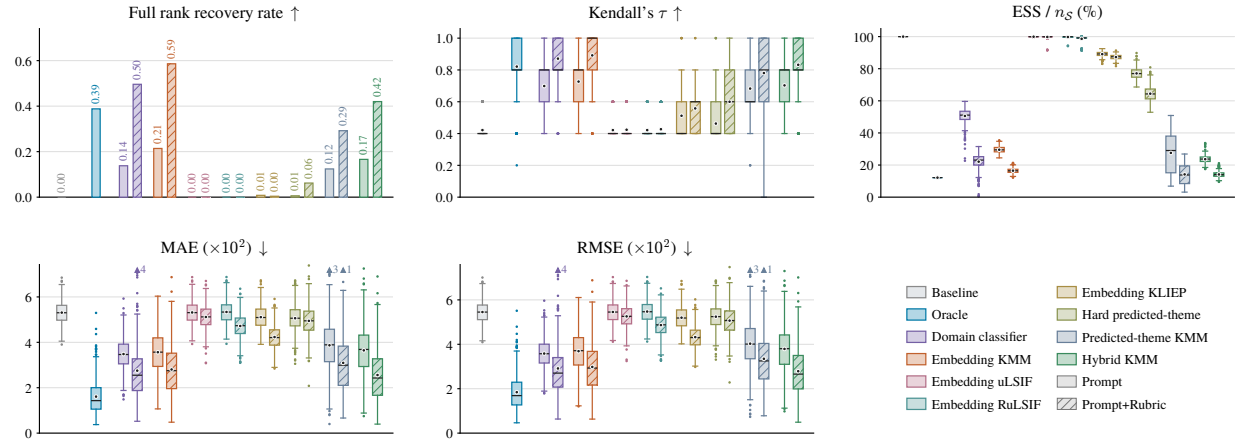

Figure 4: Comparison of reweighting methods under Case 1 using PC30 representations.

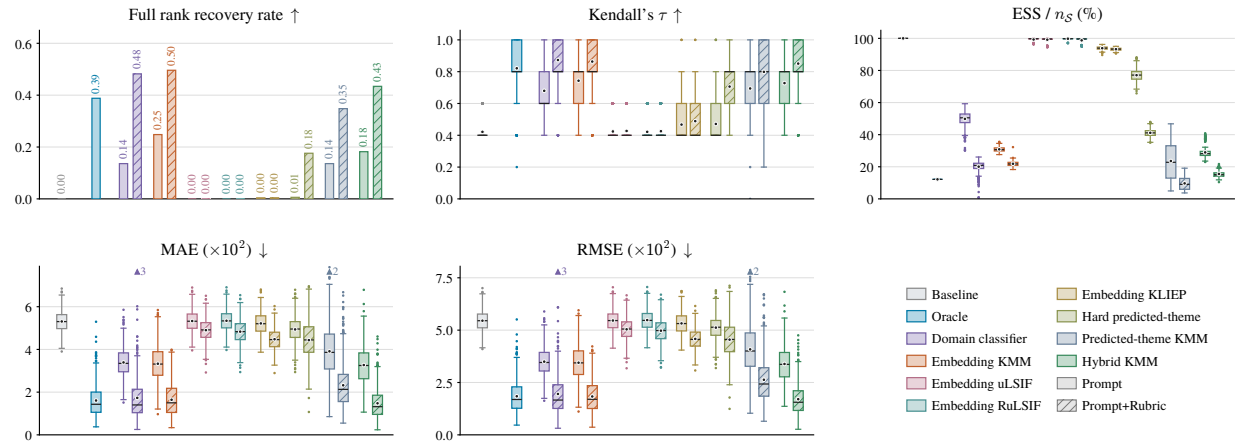

Figure 5: Comparison of reweighting methods under Case 1 using PC50 representations.

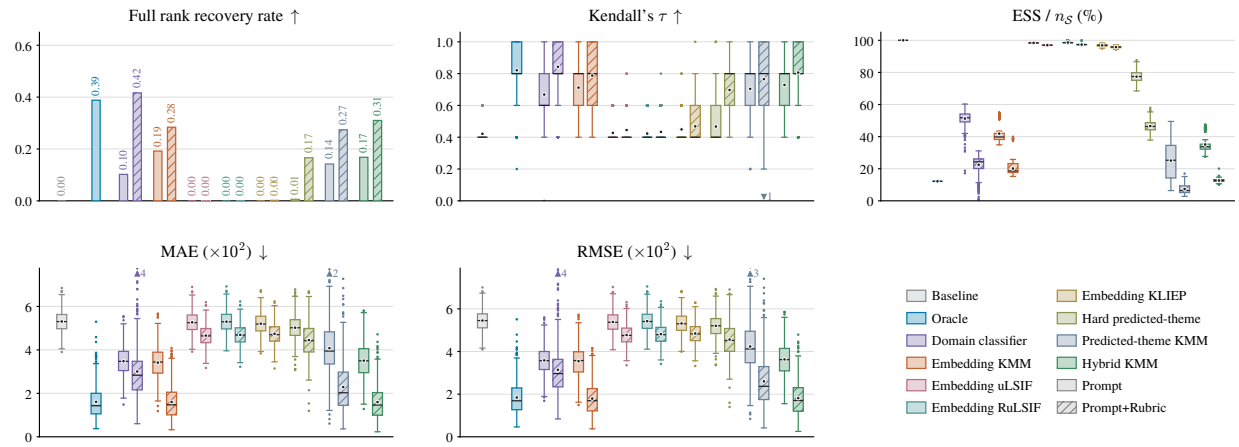

Figure 6: Comparison of reweighting methods under Case 1 using PC80 representations.

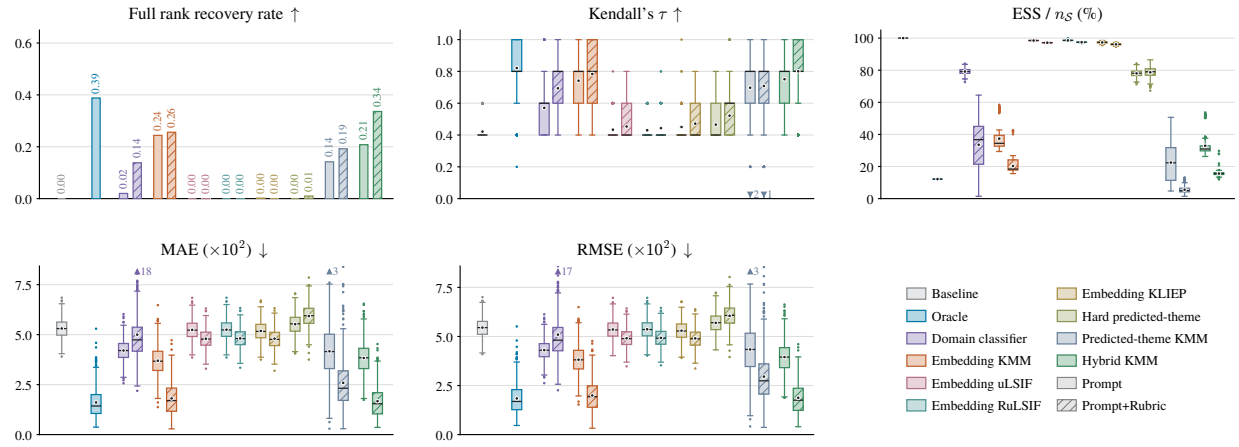

Figure 7: Comparison of reweighting methods under Case 1 using PC95 representations.

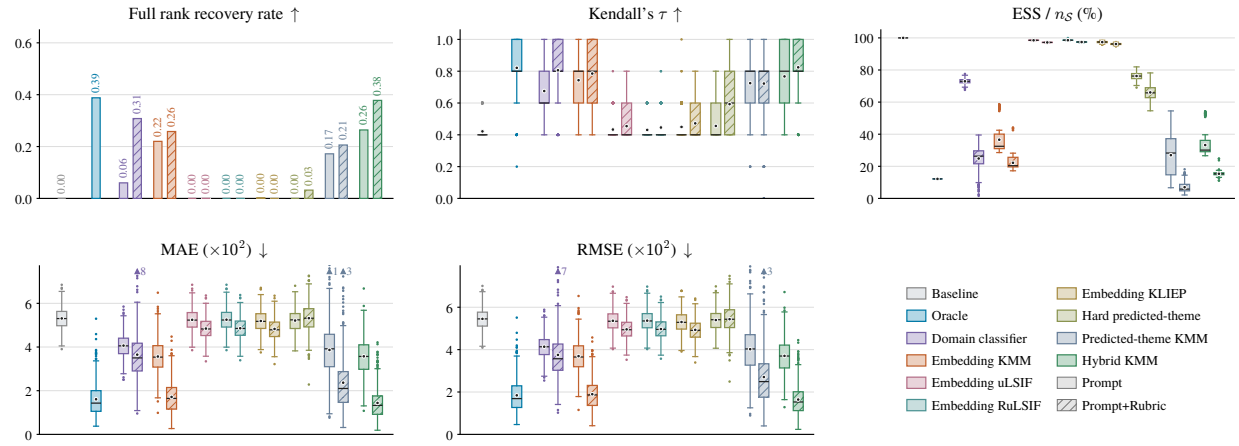

Figure 8: Comparison of reweighting methods under Case 1 using full embeddings.

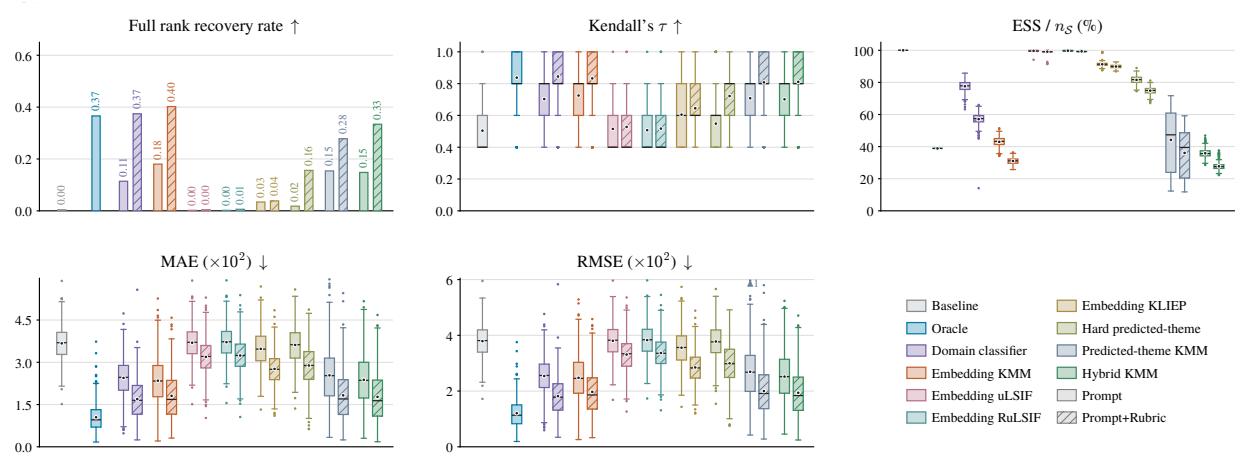

Figure 9: Comparison of reweighting methods under Case 2 using PC30 representations.

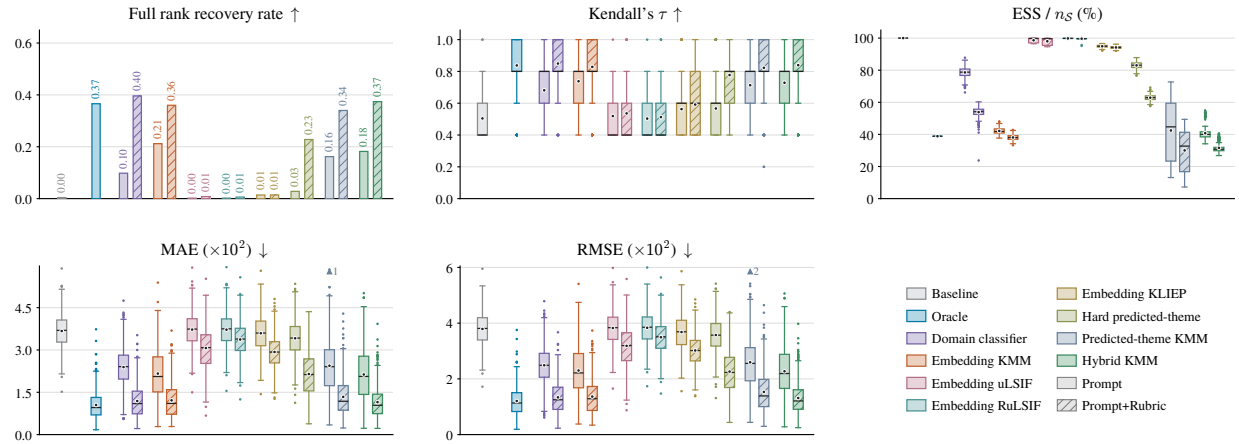

Figure 10: Comparison of reweighting methods under Case 2 using PC50 representations.

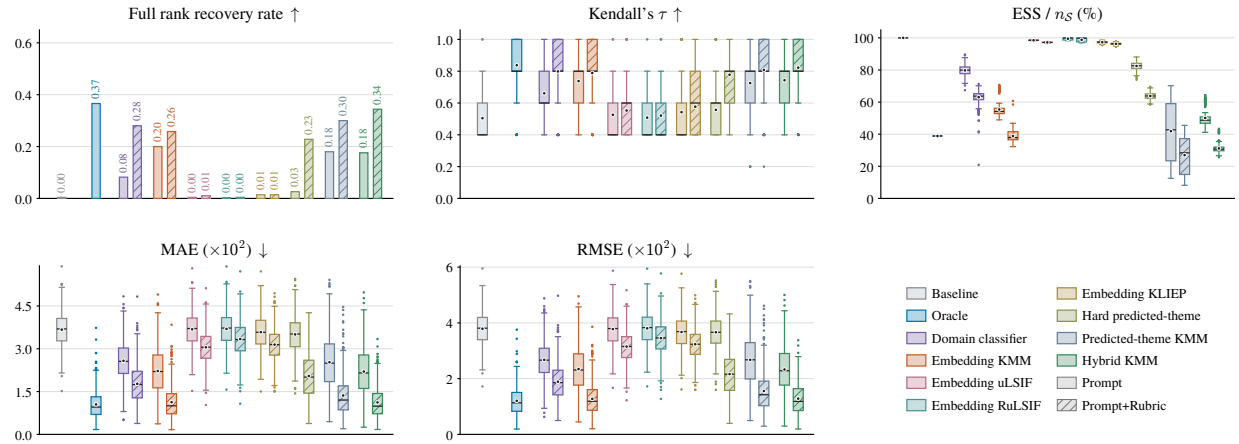

Figure 11: Comparison of reweighting methods under Case 2 using PC80 representations.

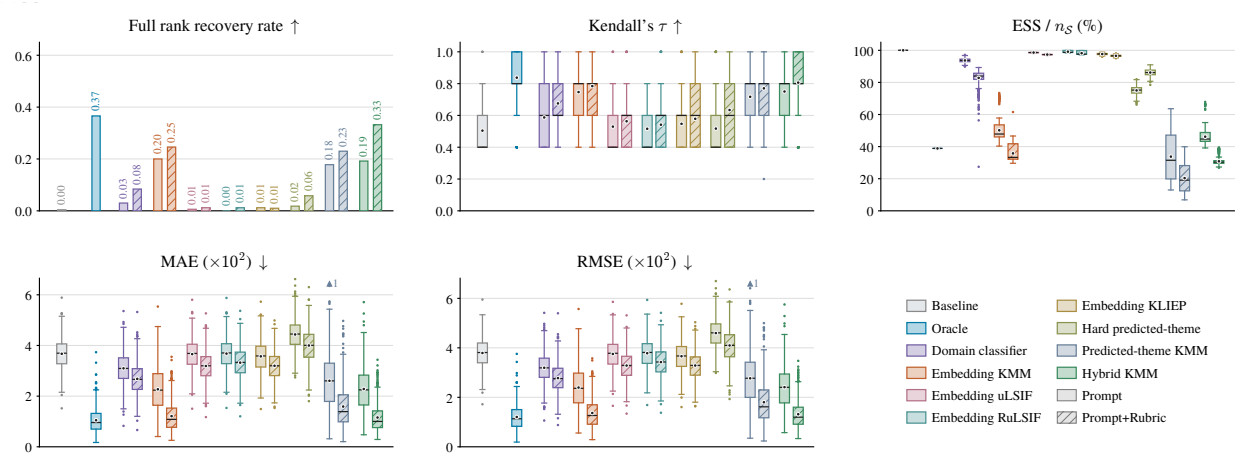

Figure 12: Comparison of reweighting methods under Case 2 using PC95 representations.

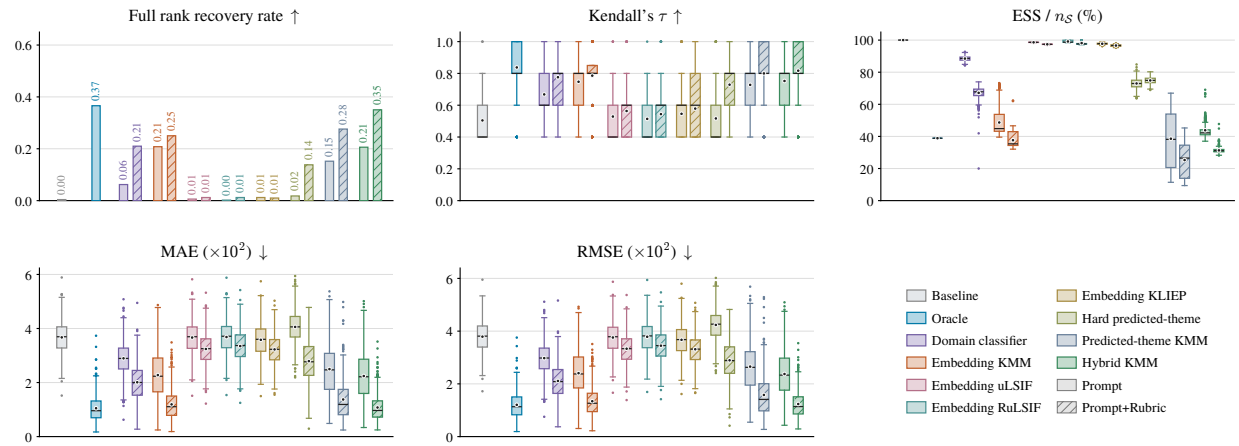

Figure 13: Comparison of reweighting methods under Case 2 using full embeddings.

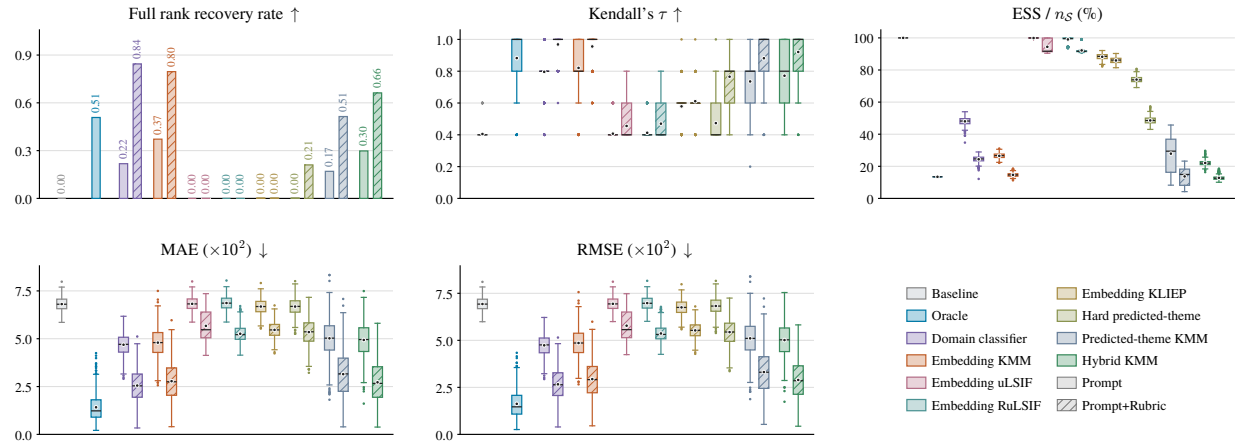

Figure 14: Comparison of reweighting methods under Case 3 using PC30 representations.

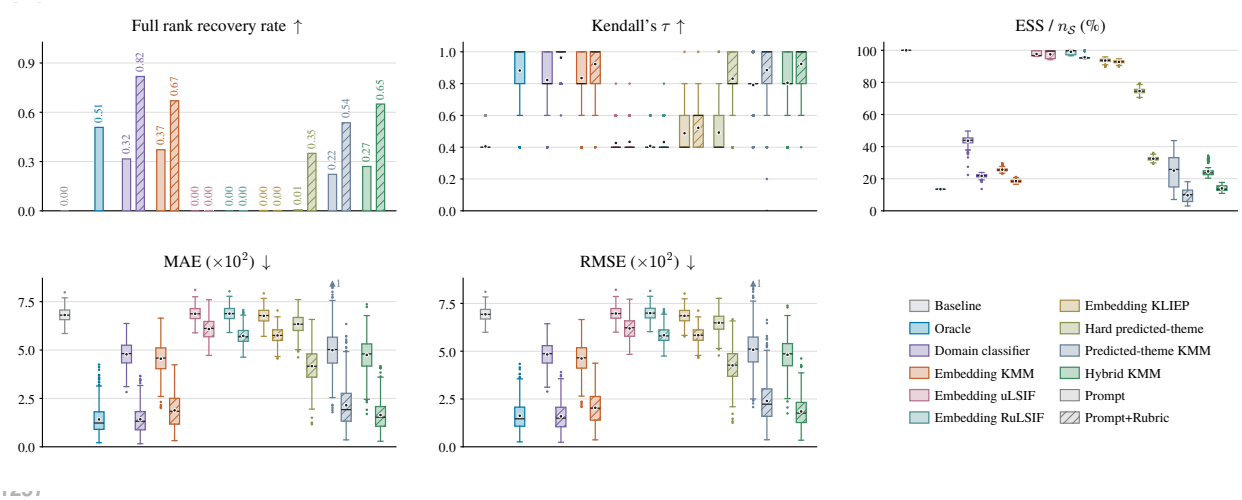

Figure 15: Comparison of reweighting methods under Case 3 using PC50 representations.

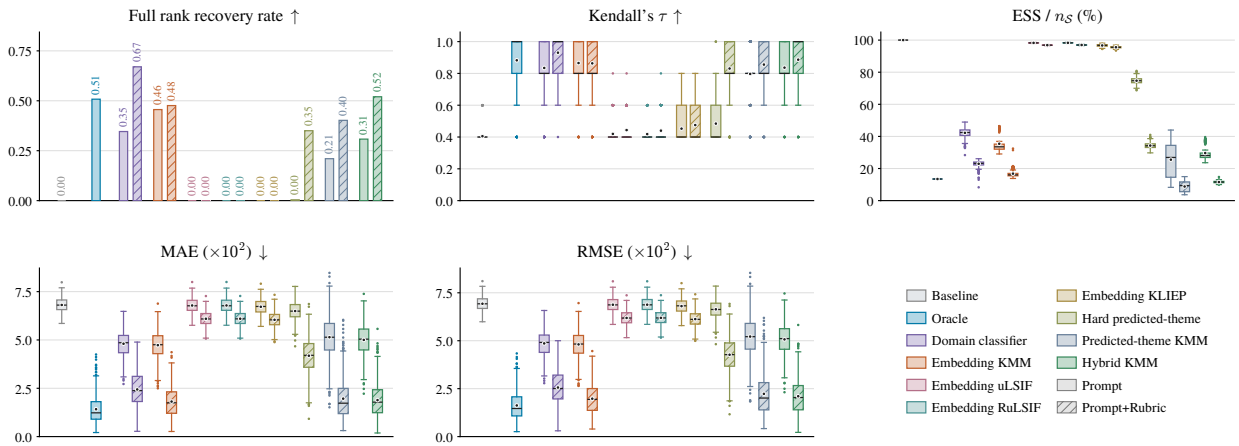

Figure 16: Comparison of reweighting methods under Case 3 using PC80 representations.

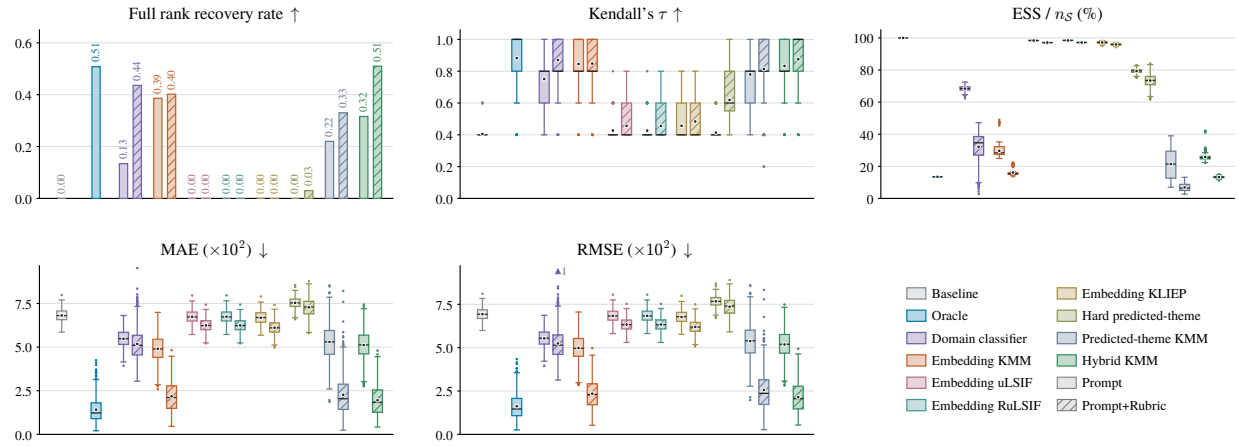

Figure 17: Comparison of reweighting methods under Case 3 using PC95 representations.

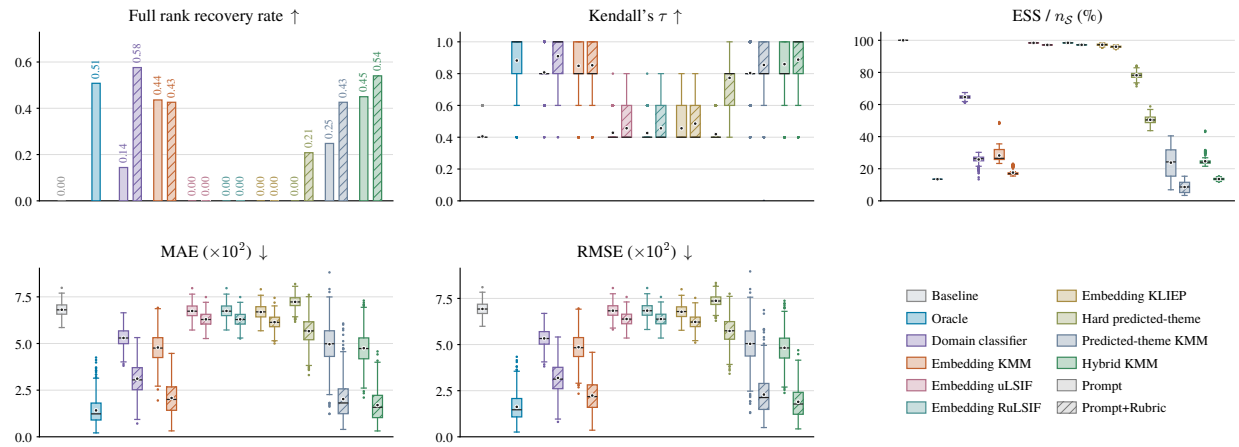

Figure 18: Comparison of reweighting methods under Case 3 using full embeddings.

B Covariate-Adaptive Fusion Estimator: Detailed Algorithm and Theoretical Guarantees

We give the detailed construction of CAFE for the different-evaluator setting and establish conditions for efficient inference. The same-evaluator specialization is given in Appendix C.

B.1 Algorithm Details

For a fixed $L \geq 2$, we randomly partition the source and target samples into L equal-sized folds, independently of the observed data. For each group $g \in \{\mathcal{S}, \mathcal{T}\}$, let $\mathcal{I}_{g,\ell}$ denote the indices in evaluation fold ℓ, and let $\mathcal{I}^c_{g,\ell} = \{1, \ldots, n_g\} \setminus \mathcal{I}_{g,\ell}$ denote the corresponding training indices. Define $n_{g,\ell} = |\mathcal{I}_{g,\ell}|$ and $n_{g,-\ell} = |\mathcal{I}^c_{g,\ell}|$, and let $\mathcal{D}_{-\ell}$ collect the training observations from both groups. The empirical averages over the evaluation and training folds are

$$\mathbb{P}_{g,\ell} f = \frac{1}{|\mathcal{I}_{g,\ell}|} \sum_{i \in \mathcal{I}_{g,\ell}} f(O^g_i), \qquad \mathbb{P}_{g,-\ell} f = \frac{1}{|\mathcal{I}^c_{g,\ell}|} \sum_{i \in \mathcal{I}^c_{g,\ell}} f(O^g_i).$$

For nuisance estimation, we first estimate the density ratio q using all source and target covariates and hold $\widehat{q}$ fixed across folds. For each fold ℓ, we estimate the target evaluation probability from the corresponding training observations as $\widehat{\alpha}_\ell = \mathbb{P}_{\mathcal{T},-\ell} R = n^{-1}_{\mathcal{T},-\ell} \sum_{j \in \mathcal{I}^c_{\mathcal{T},\ell}} R_j$ and fit m and $\widetilde{m}$ on $\mathcal{D}_{-\ell}$, yielding $\widehat{m}_\ell$ and $\widehat{\widetilde{m}}_\ell$. Given these estimates, define

$$\widehat{Q}_{-\ell}(a) = \frac{1}{2} \mathbb{P}_{\mathcal{T},-\ell} \left[\frac{R}{\widehat{\alpha}_\ell} \{\widetilde{Y} - \widehat{\widetilde{m}}_\ell(X) - a(X)\{Y - \widehat{m}_\ell(X)\}\}^2 \right] + \frac{\widehat{\alpha}_\ell \pi_{\mathcal{T}}}{2\pi_{\mathcal{S}}} \mathbb{P}_{\mathcal{S},-\ell} \left[\widehat{q}(X)^2 a(X)^2 \{Y - \widehat{m}_\ell(X)\}^2 \right]. \tag{6}$$

We estimate a on each training fold by

$$\widehat{a}_\ell \in \arg\min_{a \in \mathcal{H}} \left\{ \widehat{Q}_{-\ell}(a) + \frac{\lambda_{a,\ell}}{2} \|a\|^2_{\mathcal{H}} \right\}. \tag{7}$$

The corresponding source coefficient is $\widehat{w}_\ell(X) = \widehat{q}(X)\widehat{a}_\ell(X)$.

Using the fold-specific estimates $(\widehat{m}_\ell, \widehat{\widetilde{m}}_\ell, \widehat{a}_\ell, \widehat{\alpha}_\ell)$, we evaluate the EIF contributions on the held-out fold ℓ. For $j \in \mathcal{I}_{\mathcal{T},\ell}$ and $i \in \mathcal{I}_{\mathcal{S},\ell}$, define

$$\widehat{\mathcal{E}}^{\mathcal{T}}_j = \widehat{\widetilde{m}}_\ell(X_j) + \frac{R_j}{\widehat{\alpha}_\ell} \left[\widetilde{Y}_j - \widehat{\widetilde{m}}_\ell(X_j) - \widehat{a}_\ell(X_j)\{Y_j - \widehat{m}_\ell(X_j)\} \right], \tag{8}$$

$$\widehat{\mathcal{E}}^{\mathcal{S}}_i = \widehat{q}(X_i)\widehat{a}_\ell(X_i)\{Y_i - \widehat{m}_\ell(X_i)\}. \tag{9}$$

The fold-specific estimator is

$$\widehat{\mu}_\ell = \mathbb{P}_{\mathcal{T},\ell}\widehat{\mathcal{E}}^{\mathcal{T}} + \mathbb{P}_{\mathcal{S},\ell}\widehat{\mathcal{E}}^{\mathcal{S}}, \tag{10}$$

and the final cross-fitted estimator is $\widehat{\mu} = L^{-1}\sum_{\ell=1}^{L}\widehat{\mu}_\ell$.

We now describe variance estimation using the same held-out contributions. Let $N = n_{\mathcal{S}} + n_{\mathcal{T}}$ and $\pi_g = n_g/N$. The asymptotic variance of $\sqrt{N}(\widehat{\mu} - \mu)$ is estimated by

$$\widehat{V} = \frac{1}{\pi_{\mathcal{T}} n_{\mathcal{T}}}\sum_{j=1}^{n_{\mathcal{T}}}(\widehat{\mathcal{E}}_j^{\mathcal{T}} - \widehat{\mu})^2 + \frac{1}{\pi_{\mathcal{S}} n_{\mathcal{S}}}\sum_{i=1}^{n_{\mathcal{S}}}(\widehat{\mathcal{E}}_i^{\mathcal{S}})^2. \tag{11}$$

The target contribution is centered at $\widehat{\mu}$, whereas the source contribution is centered at zero, matching their population means. Thus, the estimated variance of $\widehat{\mu}$ is $\widehat{V}/N$, with standard error $\sqrt{\widehat{V}/N}$. Algorithm 2 summarizes the full procedure, and Theorem 2 establishes its asymptotic validity.

Algorithm 2 Covariate-Adaptive Fusion Estimator (CAFE) in the Different-Evaluator Setting

Input: Source $\mathcal{D}_{\mathcal{S}} = \{(X_i^{\mathcal{S}}, Y_i^{\mathcal{S}})\}_{i=1}^{n_{\mathcal{S}}}$, target $\mathcal{D}_{\mathcal{T}} = \{(X_j^{\mathcal{T}}, R_j, R_j Y_j^{\mathcal{T}}, R_j \widetilde{Y}_j^{\mathcal{T}})\}_{j=1}^{n_{\mathcal{T}}}$; fold count L; kernel K; positive penalties $\{\lambda_{a,\ell}\}_{\ell=1}^{L}$; learners $\mathcal{A}_m, \mathcal{A}_{\widetilde{m}}$; density-ratio learner $\mathcal{A}_q$ for $q = p_{\mathcal{T}}/p_{\mathcal{S}}$.

Output: Estimate $\widehat{\mu}$ and variance estimate $\widehat{V}$.

$N \leftarrow n_{\mathcal{S}} + n_{\mathcal{T}}$, $\pi_g \leftarrow n_g/N$, for $g \in \{\mathcal{S}, \mathcal{T}\}$.
Fit $\widehat{q}$ using $\mathcal{A}_q$ on all source and target covariates in $(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}})$; keep it fixed across folds.
Randomly partition each sample into a fixed number $L \geq 2$ of equal folds: $\{1, \ldots, n_g\} = \bigsqcup_{\ell=1}^{L}\mathcal{I}_{g,\ell}$.
for $\ell = 1, \ldots, L$ **do**
  $\mathcal{I}_{g,\ell}^c \leftarrow \{1, \ldots, n_g\} \setminus \mathcal{I}_{g,\ell}$, $g \in \{\mathcal{S}, \mathcal{T}\}$.
  $\mathcal{D}_{\mathcal{S},-\ell} \leftarrow \{(X_i^{\mathcal{S}}, Y_i^{\mathcal{S}}) : i \in \mathcal{I}_{\mathcal{S},\ell}^c\}$.
  $\mathcal{D}_{\mathcal{T},-\ell} \leftarrow \{(X_j^{\mathcal{T}}, R_j, R_j Y_j^{\mathcal{T}}, R_j \widetilde{Y}_j^{\mathcal{T}}) : j \in \mathcal{I}_{\mathcal{T},\ell}^c\}$.
  $\mathcal{D}_{-\ell} \leftarrow (\mathcal{D}_{\mathcal{S},-\ell}, \mathcal{D}_{\mathcal{T},-\ell})$.
  $\widehat{\alpha}_\ell \leftarrow \mathbb{P}_{\mathcal{T},-\ell}R = \dfrac{1}{n_{\mathcal{T},-\ell}}\displaystyle\sum_{j \in \mathcal{I}_{\mathcal{T},\ell}^c} R_j$.
  $\widehat{m}_\ell \leftarrow \mathcal{A}_m(\mathcal{D}_{-\ell})$, $\widehat{\widetilde{m}}_\ell \leftarrow \mathcal{A}_{\widetilde{m}}(\mathcal{D}_{-\ell})$.
  Form $\widehat{Q}_{-\ell}(a)$ in (6) on $\mathcal{D}_{-\ell}$ using $(\widehat{m}_\ell, \widehat{\widetilde{m}}_\ell, \widehat{q}, \widehat{\alpha}_\ell)$.
  $\widehat{a}_\ell \leftarrow \arg\min_{a \in \mathcal{H}}\left\{\widehat{Q}_{-\ell}(a) + \dfrac{\lambda_{a,\ell}}{2}\|a\|_{\mathcal{H}}^2\right\}$.
  $\widehat{w}_\ell(X_i^{\mathcal{S}}) \leftarrow \widehat{a}_\ell(X_i^{\mathcal{S}})\widehat{q}(X_i^{\mathcal{S}})$, $i = 1, \ldots, n_{\mathcal{S}}$.
  For $j \in \mathcal{I}_{\mathcal{T},\ell}$, set $\widehat{\mathcal{E}}_j^{\mathcal{T}} \leftarrow \widehat{\widetilde{m}}_\ell(X_j^{\mathcal{T}}) + \dfrac{R_j}{\widehat{\alpha}_\ell}\{\widetilde{Y}_j^{\mathcal{T}} - \widehat{\widetilde{m}}_\ell(X_j^{\mathcal{T}}) - \widehat{a}_\ell(X_j^{\mathcal{T}})[Y_j^{\mathcal{T}} - \widehat{m}_\ell(X_j^{\mathcal{T}})]\}$.
  $\widehat{\mathcal{E}}_i^{\mathcal{S}} \leftarrow \widehat{w}_\ell(X_i^{\mathcal{S}})\{Y_i^{\mathcal{S}} - \widehat{m}_\ell(X_i^{\mathcal{S}})\}$, $i \in \mathcal{I}_{\mathcal{S},\ell}$.
end for
$\overline{\mathcal{E}}_g \leftarrow \dfrac{1}{n_g}\displaystyle\sum_{i=1}^{n_g}\widehat{\mathcal{E}}_i^{g}$, $g \in \{\mathcal{S}, \mathcal{T}\}$.
$\widehat{\mu} \leftarrow \overline{\mathcal{E}}_{\mathcal{T}} + \overline{\mathcal{E}}_{\mathcal{S}}$.
$\widehat{V} \leftarrow \dfrac{1}{\pi_{\mathcal{T}} n_{\mathcal{T}}}\displaystyle\sum_{j=1}^{n_{\mathcal{T}}}(\widehat{\mathcal{E}}_j^{\mathcal{T}} - \widehat{\mu})^2 + \dfrac{1}{\pi_{\mathcal{S}} n_{\mathcal{S}}}\sum_{i=1}^{n_{\mathcal{S}}}(\widehat{\mathcal{E}}_i^{\mathcal{S}})^2$.
return $\widehat{\mu}, \widehat{V}$.

B.2 Theoretical Guarantees

We now give sufficient conditions for asymptotic normality and efficiency of $\widehat{\mu}$. Throughout, L is fixed. Write $\|f\|_{2,g} = \{\mathbb{E}_g(f^2)\}^{1/2}$, with fitted functions held fixed inside population expectations, and let $\pi_g^0 \in (0,1)$ denote the limiting sample proportions.

Assumption 2 (Regularity and boundedness)**.** *There exist constants $M, K_0 < \infty$ such that $\|Y\|_{\infty,g} \leq M$ for $g \in \{\mathcal{S}, \mathcal{T}\}$, $\|\widetilde{Y}\|_{\infty,\mathcal{T}} \leq M$, and $\sup_{x \in \mathcal{X}}\mathsf{K}(x,x) \leq K_0$. The RKHS $\mathcal{H}$ is separable with a fixed measurable kernel. There exists a uniformly bounded $P_{\mathcal{S}}$-Donsker class $\mathcal{Q}$ of nonnegative functions containing q such that $\Pr(\widehat{q} \in \mathcal{Q}) \to 1$. Moreover, $\max_{\ell \leq L}\|\widehat{m}_\ell\|_{\infty,\mathcal{S}} = O_P(1)$*

and $\max_{\ell \le L} \|\widehat{a}_\ell\|_{\mathcal{H}} = O_P(1)$. *At the limiting design proportions,*

$$\pi^0_{\mathcal{S}} p_{\mathcal{S}}(x)\sigma^2_{\mathcal{T},Y}(x) + \alpha\pi^0_{\mathcal{T}} p_{\mathcal{T}}(x)\sigma^2_{\mathcal{S},Y}(x) > 0$$

for almost every $x \in \mathcal{X}$.

Assumption 3 (Nuisance consistency)**.** *The nuisance estimators satisfy*

$$\|\widehat{q} - q\|_{2,\mathcal{S}} + \max_{\ell \le L} \left\{ \|\widehat{\widetilde{m}}_\ell - \widetilde{m}\|_{2,\mathcal{T}} + \|\widehat{m}_\ell - m\|_{2,\mathcal{T}} + \|\widehat{m}_\ell - m\|_{2,\mathcal{S}} \right\} = o_P(1),$$

and

$$\max_{\ell \le L} \left\{ \mathbb{P}_{\mathcal{T},\ell}[(\widehat{a}_\ell - a^*)^2] + \mathbb{P}_{\mathcal{S},\ell}[(\widehat{w}_\ell - qa^*)^2] \right\} = o_P(1).$$

Assumption 4 (Product rate)**.**

$$\|\widehat{q} - q\|_{2,\mathcal{S}} \max_{\ell \le L} \|\widehat{m}_\ell - m\|_{2,\mathcal{S}} = o_P(N^{-1/2}).$$

Under these conditions, we obtain the following asymptotic result. The proof is provided in Appendix D.3.

Theorem 2. *Under Assumption 1, further suppose Assumptions 2–4 hold. Let* $n_{\mathcal{S}}, n_{\mathcal{T}} \to \infty$ *with* $\pi_g = \pi^0_g + o(1)$ *and* $\pi^0_g \in (0,1)$ *for* $g \in \{\mathcal{S}, \mathcal{T}\}$. *Then* $V = \lim_{N\to\infty} \mathrm{Var}\{\phi_2(O)\}$ *exists and*

$$\sqrt{N}(\widehat{\mu} - \mu) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \phi_2(O_i) + o_P(1) \xrightarrow{d} \mathcal{N}(0, V). \tag{12}$$

The estimator is semiparametrically efficient at the limiting design proportions $(\pi^0_{\mathcal{S}}, \pi^0_{\mathcal{T}})$. *Moreover, the variance estimator* $\widehat{V}$ *in* (11) *is consistent for* V.

Remark 2. Assumption 2 imposes standard boundedness conditions and a Donsker condition on the fitted density ratio. Assumption 3 requires consistency of the nuisance estimators but does not impose parametric convergence rates on the cross-fitted regression and coefficient estimators, allowing flexible nuisance estimation. Assumption 4 requires the density-ratio and source-regression errors to satisfy a product rate in $L_2(P_{\mathcal{S}})$. A simple sufficient condition is

$$\|\widehat{q} - q\|_{2,\mathcal{S}} = o_P(N^{-1/4}), \qquad \max_{\ell \le L} \|\widehat{m}_\ell - m\|_{2,\mathcal{S}} = o_P(N^{-1/4}),$$

while the product condition also permits asymmetric rates when one nuisance estimator converges faster than the other.

Remark 3. When target scores are scarce, samples can be partitioned within the source, evaluated-target, and unevaluated-target strata to distribute evaluated target observations evenly across folds. Theorem 2 continues to hold for uniformly randomized, balanced partitions within these strata under Assumptions 2–4.

C Theoretical Details of the Same-Evaluator Setting

The same-evaluator setting is obtained from the different-evaluator model by setting $Y = \widetilde{Y}$ and $m = \widetilde{m}$. A further simplification occurs when the full conditional outcome distribution is shared across populations. In this case, $p_{\mathcal{S}}(y \mid x) = p_{\mathcal{T}}(y \mid x)$ implies $\sigma^2_{\mathcal{S},Y}(x) = \sigma^2_{\mathcal{T},Y}(x)$, yielding the distribution-aligned fusion IF studied in related data-fusion settings (Li and Luedtke, 2023; Qiu et al., 2024; Li et al., 2025a).

Corollary 1. *Assume* $p_{\mathcal{S}}(y \mid x) = p_{\mathcal{T}}(y \mid x)$. *The EIF of* μ *is*

$$\phi_3(O) = \frac{\mathbb{I}(G = \mathcal{T})}{\pi_{\mathcal{T}}}\{m(X) - \mu\} + \frac{R p_{\mathcal{T}}(X)}{\pi_{\mathcal{S}} p_{\mathcal{S}}(X) + \alpha\pi_{\mathcal{T}} p_{\mathcal{T}}(X)}\{Y - m(X)\}.$$

Therefore, Algorithm 2 for the different-evaluator setting can be specialized to these restricted settings. In particular, under the mean-invariance assumption, Algorithm 3 adapts it to the same-evaluator setting by fitting a single regression, $\widehat{m}_\ell = \widehat{\widetilde{m}}_\ell$, to the source and evaluated-target outcomes in each training fold. The density-ratio estimate $\widehat{q}$ is shared across folds, and $\widehat{\alpha}_\ell = \mathbb{P}_{\mathcal{T},-\ell} R$

Algorithm 3 Covariate-Adaptive Fusion Estimator (CAFE) under a common conditional mean

Input: $\mathcal{D}_{\mathcal{S}} = \{(X_i^{\mathcal{S}}, Y_i^{\mathcal{S}})\}_{i=1}^{n_{\mathcal{S}}}$, $\mathcal{D}_{\mathcal{T}} = \{(X_j^{\mathcal{T}}, R_j, R_j Y_j^{\mathcal{T}})\}_{j=1}^{n_{\mathcal{T}}}$; fold count L; kernel K; positive penalties $\{\lambda_{a,\ell}\}_{\ell=1}^{L}$; common-mean learner $\mathcal{A}_m$; density-ratio learner $\mathcal{A}_q$ for $q = p_{\mathcal{T}}/p_{\mathcal{S}}$.

Output: Estimate $\widehat{\mu}$ and variance estimate $\widehat{V}$.

$N \leftarrow n_{\mathcal{S}} + n_{\mathcal{T}}, \quad \pi_g \leftarrow n_g/N, \quad g \in \{\mathcal{S}, \mathcal{T}\}$.

Fit $\widehat{q}$ using $\mathcal{A}_q$ on all source and target covariates in $(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}})$; keep it fixed across folds.

Construct folds $\mathcal{I}_{g,\ell}$ and complements $\mathcal{I}_{g,\ell}^c$ as in Algorithm 2.

for $\ell = 1, \ldots, L$ **do**

  $\mathcal{D}_{\mathcal{S},-\ell} \leftarrow \{(X_i^{\mathcal{S}}, Y_i^{\mathcal{S}}) : i \in \mathcal{I}_{\mathcal{S},\ell}^c\}$.

  $\mathcal{D}_{\mathcal{T},-\ell} \leftarrow \{(X_j^{\mathcal{T}}, R_j, R_j Y_j^{\mathcal{T}}) : j \in \mathcal{I}_{\mathcal{T},\ell}^c\}$.

  $\mathcal{D}_{-\ell} \leftarrow (\mathcal{D}_{\mathcal{S},-\ell}, \mathcal{D}_{\mathcal{T},-\ell}), \quad \widehat{m}_\ell \leftarrow \mathcal{A}_m(\mathcal{D}_{-\ell})$.

  $\widehat{\alpha}_\ell \leftarrow n_{\mathcal{T},-\ell}^{-1} \sum_{j \in \mathcal{I}_{\mathcal{T},\ell}^c} R_j$.

  Form $\widehat{Q}_{-\ell}^{(1)}(a)$ in (13) on $\mathcal{D}_{-\ell}$ using the shared regression $\widehat{m}_\ell$ and density ratio $\widehat{q}$.

  $\widehat{a}_\ell \leftarrow \arg\min_{a \in \mathcal{H}} \left\{ \widehat{Q}_{-\ell}^{(1)}(a) + \frac{\lambda_{a,\ell}}{2} \|a\|_{\mathcal{H}}^2 \right\}, \quad \widehat{w}_\ell(X_i^{\mathcal{S}}) \leftarrow \widehat{a}_\ell(X_i^{\mathcal{S}}) \widehat{q}(X_i^{\mathcal{S}}), i = 1, \ldots, n_{\mathcal{S}}$.

  $\widehat{\mathcal{E}}_j^{\mathcal{T}} \leftarrow \widehat{m}_\ell(X_j^{\mathcal{T}}) + \frac{R_j}{\widehat{\alpha}_\ell} \{1 - \widehat{a}_\ell(X_j^{\mathcal{T}})\} \{Y_j^{\mathcal{T}} - \widehat{m}_\ell(X_j^{\mathcal{T}})\}, \quad j \in \mathcal{I}_{\mathcal{T},\ell}$.

  $\widehat{\mathcal{E}}_i^{\mathcal{S}} \leftarrow \widehat{w}_\ell(X_i^{\mathcal{S}}) \{Y_i^{\mathcal{S}} - \widehat{m}_\ell(X_i^{\mathcal{S}})\}, \quad i \in \mathcal{I}_{\mathcal{S},\ell}$.

end for

$\overline{\mathcal{E}}_g \leftarrow \frac{1}{n_g} \sum_{i=1}^{n_g} \widehat{\mathcal{E}}_i^g, \quad g \in \{\mathcal{S}, \mathcal{T}\}$.

$\widehat{\mu} \leftarrow \overline{\mathcal{E}}_{\mathcal{T}} + \overline{\mathcal{E}}_{\mathcal{S}}$.

$\widehat{V} \leftarrow \frac{1}{\pi_{\mathcal{T}} n_{\mathcal{T}}} \sum_{j=1}^{n_{\mathcal{T}}} (\widehat{\mathcal{E}}_j^{\mathcal{T}} - \widehat{\mu})^2 + \frac{1}{\pi_{\mathcal{S}} n_{\mathcal{S}}} \sum_{i=1}^{n_{\mathcal{S}}} (\widehat{\mathcal{E}}_i^{\mathcal{S}})^2$.

return $\widehat{\mu}, \widehat{V}$.

is estimated as in Appendix B. The coefficient loss becomes

$$\begin{aligned} \widehat{Q}_{-\ell}^{(1)}(a) = \frac{1}{2} \mathbb{P}_{\mathcal{T},-\ell} \left[\frac{R}{\widehat{\alpha}_\ell} (1 - a(X))^2 \{Y - \widehat{m}_\ell(X)\}^2 \right] \\ + \frac{\widehat{\alpha}_\ell \pi_{\mathcal{T}}}{2\pi_{\mathcal{S}}} \mathbb{P}_{\mathcal{S},-\ell} \left[\widehat{q}(X)^2 a(X)^2 \{Y - \widehat{m}_\ell(X)\}^2 \right]. \end{aligned} \tag{13}$$

Given $\widehat{a}_\ell$ and $\widehat{w}_\ell = \widehat{q}\,\widehat{a}_\ell$, the fold-specific estimate is

$$\begin{aligned} \widehat{\mu}_\ell = \mathbb{P}_{\mathcal{T},\ell} \left[\widehat{m}_\ell(X) + \frac{R}{\widehat{\alpha}_\ell} \{1 - \widehat{a}_\ell(X)\} \{Y - \widehat{m}_\ell(X)\} \right] \\ + \mathbb{P}_{\mathcal{S},\ell} [\widehat{w}_\ell(X) \{Y - \widehat{m}_\ell(X)\}]. \end{aligned} \tag{14}$$

Under the assumptions of Theorem 2, specialized to $Y = \widetilde{Y}$, $m = \widetilde{m}$, and $\widehat{m}_\ell = \widehat{\widetilde{m}}_\ell$, the estimator in Algorithm 3 is asymptotically linear with influence function (1), evaluated at the limiting design proportions, and the corresponding variance estimator is consistent. We therefore omit the separate proof.

D PROOFS

D.1 PROOF OF PROPOSITION 1: SAME-EVALUATOR SETTING

Consider a regular one-dimensional parametric submodel indexed by $t \in (-\epsilon, \epsilon)$ with bounded score components, of the form

$$p_{g,t}(y \mid x) = p_g(y \mid x)\{1 + t\zeta_g(x, y)\}, \quad \mathbb{E}_g\{\zeta_g(X, Y) \mid X\} = 0,$$
$$p_{g,t}(x) = p_g(x)\{1 + t\kappa_g(x)\}, \quad \mathbb{E}_g\{\kappa_g(X)\} = 0, \quad g \in \{\mathcal{S}, \mathcal{T}\}.$$

Under the fixed-sample-size formulation, G indexes sample membership and does not contribute an additional likelihood factor. For the observed record $O = (G, X, R, RY)$, the observed-data likelihood contribution under the submodel P_t is

$$p_t(O \mid G) = \{p_{\mathcal{S},t}(X)p_{\mathcal{S},t}(Y \mid X)\}^{\mathbb{I}(G=\mathcal{S})} \left\{p_{\mathcal{T},t}(X)\alpha^R(1-\alpha)^{1-R}p_{\mathcal{T},t}(Y \mid X)^R\right\}^{\mathbb{I}(G=\mathcal{T})}.$$

Holding α fixed along the submodel and differentiating the log likelihood gives the observed-data score

$$s(O) = \mathbb{I}(G = \mathcal{S})\{\kappa_{\mathcal{S}}(X) + \zeta_{\mathcal{S}}(X, Y)\} + \mathbb{I}(G = \mathcal{T})\{\kappa_{\mathcal{T}}(X) + R\zeta_{\mathcal{T}}(X, Y)\}.$$

Along a submodel satisfying $\mathbb{E}_{\mathcal{S},t}(Y \mid X = x) = \mathbb{E}_{\mathcal{T},t}(Y \mid X = x)$,

$$\begin{aligned}\frac{d}{dt}m_t(x)\bigg|_{t=0} &= \mathbb{E}_{\mathcal{S}}[\{Y - m(X)\}\zeta_{\mathcal{S}}(X, Y) \mid X = x] \\ &= \mathbb{E}_{\mathcal{T}}[\{Y - m(X)\}\zeta_{\mathcal{T}}(X, Y) \mid X = x]. \end{aligned} \tag{15}$$

The tangent space is therefore

$$\mathscr{T} = \mathrm{cl}\left\{\begin{array}{l} s(O) : \kappa_g, \zeta_g \in L^2(P_g), \quad g \in \{\mathcal{S}, \mathcal{T}\}, \\ \mathbb{E}_{\mathcal{S}}[\{Y - m(X)\}\zeta_{\mathcal{S}}(X, Y) \mid X = x] = \mathbb{E}_{\mathcal{T}}[\{Y - m(X)\}\zeta_{\mathcal{T}}(X, Y) \mid X = x] \end{array}\right\},$$

where the score components satisfy the mean-zero conditions above and the conditional mean constraint holds $P_{\mathcal{T},X}$-almost surely. The source conditional score is unrestricted by this constraint where $q(X) = 0$; square-integrable scores are included through the displayed closure.

Since $\mu_t = \int m_t(x)p_{\mathcal{T},t}(x)\,dx$, differentiation over t gives

$$\begin{aligned}\frac{d}{dt}\mu_t\bigg|_{t=0} &= \int \left[\frac{d}{dt}m_t(x)\bigg|_{t=0} p_{\mathcal{T}}(x) + m(x)p_{\mathcal{T}}(x)\kappa_{\mathcal{T}}(x)\right] dx \\ &= \mathbb{E}_{\mathcal{T}}\left\{\frac{d}{dt}m_t(X)\bigg|_{t=0}\right\} + \mathbb{E}_{\mathcal{T}}[\{m(X) - \mu\}\kappa_{\mathcal{T}}(X)] \\ &= \mathbb{E}_{\mathcal{T}}[\{Y - m(X)\}\zeta_{\mathcal{T}}(X, Y)] + \mathbb{E}_{\mathcal{T}}[\{m(X) - \mu\}\kappa_{\mathcal{T}}(X)] \\ &= \mathbb{E}_{\mathcal{T}}[\{m(X) - \mu\}\{\kappa_{\mathcal{T}}(X) + R\zeta_{\mathcal{T}}(X, Y)\}] \\ &\quad + \mathbb{E}_{\mathcal{T}}\left[\frac{R}{\alpha}\{Y - m(X)\}\{\kappa_{\mathcal{T}}(X) + R\zeta_{\mathcal{T}}(X, Y)\}\right] \\ &= \mathbb{E}_{\mathcal{T}}\left[\left\{m(X) - \mu + \frac{R}{\alpha}\{Y - m(X)\}\right\} s(O)\right].\end{aligned}$$

Let

$$\phi_0(O) = \frac{\mathbb{I}(G = \mathcal{T})}{\pi_{\mathcal{T}}}\left[m(X) - \mu + \frac{R}{\alpha}\{Y - m(X)\}\right].$$

Since $\phi_0(O) = 0$ for source records,

$$\frac{d}{dt}\mu_t\bigg|_{t=0} = \pi_{\mathcal{S}}\mathbb{E}_{\mathcal{S}}\{\phi_0(O)s(O)\} + \pi_{\mathcal{T}}\mathbb{E}_{\mathcal{T}}\{\phi_0(O)s(O)\}.$$

Therefore $\phi_0(O)$ is an influence function of μ.

For an admissible measurable function a, define

$$h_a(O) = \left[\frac{\mathbb{I}(G=\mathcal{S})q(X)}{\pi_{\mathcal{S}}} - \frac{\mathbb{I}(G=\mathcal{T})R}{\alpha\pi_{\mathcal{T}}}\right] a(X)\{Y - m(X)\}.$$

Admissibility requires $h_a(O)$ to have finite second moments in both samples. The identity $\mathbb{E}_{\mathcal{S}}\{q(X)f(X)\} = \mathbb{E}_{\mathcal{T}}\{f(X)\}$ and (15) imply, for every score $s \in \mathscr{T}$,

$$\begin{aligned}
&\pi_{\mathcal{S}}\mathbb{E}_{\mathcal{S}}\{h_a(O)s(O)\} + \pi_{\mathcal{T}}\mathbb{E}_{\mathcal{T}}\{h_a(O)s(O)\} \\
&= \mathbb{E}_{\mathcal{S}}[q(X)a(X)\{Y - m(X)\}\zeta_{\mathcal{S}}(X,Y)] - \mathbb{E}_{\mathcal{T}}[a(X)\{Y - m(X)\}\zeta_{\mathcal{T}}(X,Y)] \\
&= \mathbb{E}_{\mathcal{T}}[a(X)\{\mathbb{E}_{\mathcal{S}}[\{Y - m(X)\}\zeta_{\mathcal{S}}(X,Y) \mid X] - \mathbb{E}_{\mathcal{T}}[\{Y - m(X)\}\zeta_{\mathcal{T}}(X,Y) \mid X]\}] = 0.
\end{aligned}$$

Thus $h_a \in \mathscr{T}^{\perp}$, using only the assumed one-sided overlap.

We seek a^* such that $\phi_0 + h_{a^*}$ is orthogonal to every admissible h_a. Using $R^2 = R$ and $\mathbb{E}_{\mathcal{T}}(R \mid X, Y) = \alpha$, this condition becomes

$$\begin{aligned}
0 &= \pi_{\mathcal{S}}\mathbb{E}_{\mathcal{S}}[\{\phi_0(O) + h_{a^*}(O)\}h_a(O)] + \pi_{\mathcal{T}}\mathbb{E}_{\mathcal{T}}[\{\phi_0(O) + h_{a^*}(O)\}h_a(O)] \\
&= \frac{1}{\alpha\pi_{\mathcal{T}}}\mathbb{E}_{\mathcal{T}}\left[a(X)\left\{-\sigma^2_{\mathcal{T},Y}(X) + a^*(X)\left(\sigma^2_{\mathcal{T},Y}(X) + \frac{\alpha\pi_{\mathcal{T}}}{\pi_{\mathcal{S}}}q(X)\sigma^2_{\mathcal{S},Y}(X)\right)\right\}\right].
\end{aligned}$$

It is satisfied by

$$a^*(X) = \frac{\pi_{\mathcal{S}}\sigma^2_{\mathcal{T},Y}(X)}{\pi_{\mathcal{S}}\sigma^2_{\mathcal{T},Y}(X) + \alpha\pi_{\mathcal{T}}q(X)\sigma^2_{\mathcal{S},Y}(X)},$$

which is the coefficient in Proposition 1 when $\widetilde{Y} = Y$. Since $0 \le a^*(X) \le 1$ on the target support, finite second moments give

$$\begin{aligned}
\pi_{\mathcal{S}}\mathbb{E}_{\mathcal{S}}\{h_{a^*}(O)^2\} + \pi_{\mathcal{T}}\mathbb{E}_{\mathcal{T}}\{h_{a^*}(O)^2\} &= \frac{\mathbb{E}_{\mathcal{T}}\{a^*(X)\sigma^2_{\mathcal{T},Y}(X)\}}{\alpha\pi_{\mathcal{T}}} \\
&\le \frac{\mathbb{E}_{\mathcal{T}}\{\sigma^2_{\mathcal{T},Y}(X)\}}{\alpha\pi_{\mathcal{T}}} < \infty.
\end{aligned}$$

Thus h_{a^*} is admissible. Adding h_{a^*} to ϕ_0 gives

$$\begin{aligned}
\phi_1(O) = \frac{\mathbb{I}(G=\mathcal{T})}{\pi_{\mathcal{T}}}\{m(X) - \mu\} \\
+ \frac{q(X)\{\mathbb{I}(G=\mathcal{S})\sigma^2_{\mathcal{T},Y}(X) + \mathbb{I}(G=\mathcal{T})R\sigma^2_{\mathcal{S},Y}(X)\}}{\pi_{\mathcal{S}}\sigma^2_{\mathcal{T},Y}(X) + \alpha\pi_{\mathcal{T}}q(X)\sigma^2_{\mathcal{S},Y}(X)}\{Y - m(X)\}.
\end{aligned}$$

To verify that $\phi_1 \in \mathscr{T}$, take

$$\begin{aligned}
\kappa_{\mathcal{S}}(X) &= 0, & \kappa_{\mathcal{T}}(X) &= \frac{m(X) - \mu}{\pi_{\mathcal{T}}}, \\
\zeta_{\mathcal{S}}(X,Y) &= \frac{q(X)a^*(X)}{\pi_{\mathcal{S}}}\{Y - m(X)\}, & \zeta_{\mathcal{T}}(X,Y) &= \frac{1 - a^*(X)}{\alpha\pi_{\mathcal{T}}}\{Y - m(X)\}.
\end{aligned}$$

These square-integrable components satisfy the required mean-zero conditions, and (15) follows from

$$\frac{q(X)a^*(X)\sigma^2_{\mathcal{S},Y}(X)}{\pi_{\mathcal{S}}} = \frac{\{1 - a^*(X)\}\sigma^2_{\mathcal{T},Y}(X)}{\alpha\pi_{\mathcal{T}}} \quad P_{\mathcal{T},X}\text{-almost surely.}$$

Hence $\phi_1 \in \mathscr{T}$. Since $\phi_1 = \phi_0 + h_{a^*}$ and $h_{a^*} \in \mathscr{T}^{\perp}$, ϕ_1 is the EIF of μ.

D.2 Proof of Proposition 1: Different-evaluator setting

Consider a regular one-dimensional parametric submodel indexed by $t \in (-\epsilon, \epsilon)$ with bounded score components, of the form

$$\begin{aligned}
p_{\mathcal{S},t}(y \mid x) &= p_{\mathcal{S}}(y \mid x)\{1 + t\zeta_{\mathcal{S}}(x,y)\}, & \mathbb{E}_{\mathcal{S}}\{\zeta_{\mathcal{S}}(X,Y) \mid X\} &= 0, \\
p_{\mathcal{T},t}(y,\widetilde{y} \mid x) &= p_{\mathcal{T}}(y,\widetilde{y} \mid x)\{1 + t\zeta_{\mathcal{T}}(x,y,\widetilde{y})\}, & \mathbb{E}_{\mathcal{T}}\{\zeta_{\mathcal{T}}(X,Y,\widetilde{Y}) \mid X\} &= 0, \\
p_{g,t}(x) &= p_g(x)\{1 + t\kappa_g(x)\}, & \mathbb{E}_g\{\kappa_g(X)\} &= 0, \quad g \in \{\mathcal{S},\mathcal{T}\}.
\end{aligned}$$

Under the fixed-sample-size formulation, G indexes sample membership and does not contribute an additional likelihood factor. For the observed record $O = (G, X, R, RY, \mathbb{I}(G = \mathcal{T})R\widetilde{Y})$, the observed-data likelihood contribution under the submodel P_t is

$$p_t(O \mid G) = \{p_{\mathcal{S},t}(X)p_{\mathcal{S},t}(Y \mid X)\}^{\mathbb{I}(G=\mathcal{S})} \left\{p_{\mathcal{T},t}(X)\alpha^R(1-\alpha)^{1-R}p_{\mathcal{T},t}(Y, \widetilde{Y} \mid X)^R\right\}^{\mathbb{I}(G=\mathcal{T})}.$$

Holding α fixed along the submodel and differentiating the log likelihood gives the observed-data score

$$s(O) = \mathbb{I}(G = \mathcal{S})\{\kappa_{\mathcal{S}}(X) + \zeta_{\mathcal{S}}(X, Y)\} + \mathbb{I}(G = \mathcal{T})\{\kappa_{\mathcal{T}}(X) + R\zeta_{\mathcal{T}}(X, Y, \widetilde{Y})\}.$$

Along a submodel satisfying $\mathbb{E}_{\mathcal{S},t}(Y \mid X = x) = \mathbb{E}_{\mathcal{T},t}(Y \mid X = x) = m_t(x)$,

$$\begin{aligned}\frac{d}{dt}m_t(x)\Big|_{t=0} &= \mathbb{E}_{\mathcal{S}}[\{Y - m(X)\}\zeta_{\mathcal{S}}(X, Y) \mid X = x] \\ &= \mathbb{E}_{\mathcal{T}}[\{Y - m(X)\}\zeta_{\mathcal{T}}(X, Y, \widetilde{Y}) \mid X = x]. \end{aligned} \tag{16}$$

The tangent space is therefore

$$\mathscr{T} = \mathrm{cl}\left\{\begin{aligned}&s(O) : \kappa_g, \zeta_g \in L^2(P_g), \quad g \in \{\mathcal{S}, \mathcal{T}\}, \\ &\mathbb{E}_{\mathcal{S}}[\{Y - m(X)\}\zeta_{\mathcal{S}}(X, Y) \mid X = x] = \mathbb{E}_{\mathcal{T}}[\{Y - m(X)\}\zeta_{\mathcal{T}}(X, Y, \widetilde{Y}) \mid X = x]\end{aligned}\right\},$$

where the score components satisfy the mean-zero conditions above and the conditional mean constraint holds $P_{\mathcal{T},X}$-almost surely. The source conditional score is unrestricted by this constraint where $q(X) = 0$; square-integrable scores are included through the displayed closure.

Since $\mu_t = \int \widetilde{m}_t(x)p_{\mathcal{T},t}(x)\,dx$, differentiation gives

$$\begin{aligned}\frac{d}{dt}\mu_t\Big|_{t=0} &= \int \left[\frac{d}{dt}\widetilde{m}_t(x)\Big|_{t=0} p_{\mathcal{T}}(x) + \widetilde{m}(x)p_{\mathcal{T}}(x)\kappa_{\mathcal{T}}(x)\right] dx \\ &= \mathbb{E}_{\mathcal{T}}\left\{\frac{d}{dt}\widetilde{m}_t(X)\Big|_{t=0}\right\} + \mathbb{E}_{\mathcal{T}}[\{\widetilde{m}(X) - \mu\}\kappa_{\mathcal{T}}(X)] \\ &= \mathbb{E}_{\mathcal{T}}[\{\widetilde{Y} - \widetilde{m}(X)\}\zeta_{\mathcal{T}}(X, Y, \widetilde{Y})] + \mathbb{E}_{\mathcal{T}}[\{\widetilde{m}(X) - \mu\}\kappa_{\mathcal{T}}(X)] \\ &= \mathbb{E}_{\mathcal{T}}[\{\widetilde{m}(X) - \mu\}\{\kappa_{\mathcal{T}}(X) + R\zeta_{\mathcal{T}}(X, Y, \widetilde{Y})\}] \\ &\quad + \mathbb{E}_{\mathcal{T}}\left[\frac{R}{\alpha}\{\widetilde{Y} - \widetilde{m}(X)\}\{\kappa_{\mathcal{T}}(X) + R\zeta_{\mathcal{T}}(X, Y, \widetilde{Y})\}\right] \\ &= \mathbb{E}_{\mathcal{T}}\left[\left\{\widetilde{m}(X) - \mu + \frac{R}{\alpha}\{\widetilde{Y} - \widetilde{m}(X)\}\right\} s(O)\right].\end{aligned}$$

Let

$$\phi_0(O) = \frac{\mathbb{I}(G = \mathcal{T})}{\pi_{\mathcal{T}}}\left[\widetilde{m}(X) - \mu + \frac{R}{\alpha}\{\widetilde{Y} - \widetilde{m}(X)\}\right].$$

Since $\phi_0(O) = 0$ for source records,

$$\frac{d}{dt}\mu_t\Big|_{t=0} = \pi_{\mathcal{S}}\mathbb{E}_{\mathcal{S}}\{\phi_0(O)s(O)\} + \pi_{\mathcal{T}}\mathbb{E}_{\mathcal{T}}\{\phi_0(O)s(O)\}.$$

Therefore $\phi_0(O)$ is an influence function of μ.

For an admissible measurable function a, define

$$h_a(O) = \left[\frac{\mathbb{I}(G = \mathcal{S})q(X)}{\pi_{\mathcal{S}}} - \frac{\mathbb{I}(G = \mathcal{T})R}{\alpha\pi_{\mathcal{T}}}\right] a(X)\{Y - m(X)\}.$$

Admissibility again requires finite second moments in both samples. Using $\mathbb{E}_{\mathcal{S}}\{q(X)f(X)\} = \mathbb{E}_{\mathcal{T}}\{f(X)\}$ and (16), we obtain

$$\begin{aligned}&\pi_{\mathcal{S}}\mathbb{E}_{\mathcal{S}}\{h_a(O)s(O)\} + \pi_{\mathcal{T}}\mathbb{E}_{\mathcal{T}}\{h_a(O)s(O)\} \\ &= \mathbb{E}_{\mathcal{S}}[q(X)a(X)\{Y - m(X)\}\zeta_{\mathcal{S}}(X, Y)] - \mathbb{E}_{\mathcal{T}}[a(X)\{Y - m(X)\}\zeta_{\mathcal{T}}(X, Y, \widetilde{Y})] \\ &= \mathbb{E}_{\mathcal{T}}\Big[a(X)\left\{\mathbb{E}_{\mathcal{S}}[\{Y - m(X)\}\zeta_{\mathcal{S}}(X, Y) \mid X] - \mathbb{E}_{\mathcal{T}}[\{Y - m(X)\}\zeta_{\mathcal{T}}(X, Y, \widetilde{Y}) \mid X]\right\}\Big] = 0.\end{aligned}$$

Thus $h_a \in \mathscr{T}^\perp$. To determine how much information to borrow from benchmark-evaluator scores, we seek a^* such that $\phi_0 + h_{a^*}$ is orthogonal to every admissible h_a. Random selection for evaluation gives

$$\begin{aligned} 0 &= \pi_{\mathcal{S}}\mathbb{E}_{\mathcal{S}}[\{\phi_0(O) + h_{a^*}(O)\}h_a(O)] + \pi_{\mathcal{T}}\mathbb{E}_{\mathcal{T}}[\{\phi_0(O) + h_{a^*}(O)\}h_a(O)] \\ &= \frac{1}{\alpha\pi_{\mathcal{T}}}\mathbb{E}_{\mathcal{T}}\left[a(X)\left\{-\gamma_{\mathcal{T}}(X) + a^*(X)\left(\sigma^2_{\mathcal{T},Y}(X) + \frac{\alpha\pi_{\mathcal{T}}}{\pi_{\mathcal{S}}}q(X)\sigma^2_{\mathcal{S},Y}(X)\right)\right\}\right]. \end{aligned}$$

This condition is satisfied by

$$a^*(X) = \frac{\pi_{\mathcal{S}}\gamma_{\mathcal{T}}(X)}{\pi_{\mathcal{S}}\sigma^2_{\mathcal{T},Y}(X) + \alpha\pi_{\mathcal{T}}q(X)\sigma^2_{\mathcal{S},Y}(X)}.$$

Conditional Cauchy–Schwarz gives $\gamma_{\mathcal{T}}(x)^2 \le \sigma^2_{\mathcal{T},\widetilde{Y}}(x)\sigma^2_{\mathcal{T},Y}(x)$ and hence $0 \le a^*(x)\gamma_{\mathcal{T}}(x) \le \sigma^2_{\mathcal{T},\widetilde{Y}}(x)$. Finite second moments then yield

$$\pi_{\mathcal{S}}\mathbb{E}_{\mathcal{S}}\{h_{a^*}(O)^2\} + \pi_{\mathcal{T}}\mathbb{E}_{\mathcal{T}}\{h_{a^*}(O)^2\} = \frac{\mathbb{E}_{\mathcal{T}}\{a^*(X)\gamma_{\mathcal{T}}(X)\}}{\alpha\pi_{\mathcal{T}}} \le \frac{\mathbb{E}_{\mathcal{T}}\{\sigma^2_{\mathcal{T},\widetilde{Y}}(X)\}}{\alpha\pi_{\mathcal{T}}} < \infty,$$

establishing that h_{a^*} is square-integrable in both samples. Adding h_{a^*} to ϕ_0 gives

$$\begin{aligned} \phi_2(O) = \frac{\mathbb{I}(G=\mathcal{T})}{\pi_{\mathcal{T}}}\{\widetilde{m}(X) - \mu\} + \frac{\mathbb{I}(G=\mathcal{S})p_{\mathcal{T}}(X)}{\pi_{\mathcal{S}}p_{\mathcal{S}}(X)}a^*(X)\{Y - m(X)\} \\ + \frac{\mathbb{I}(G=\mathcal{T})R}{\alpha\pi_{\mathcal{T}}}\left[\widetilde{Y} - \widetilde{m}(X) - a^*(X)\{Y - m(X)\}\right]. \end{aligned}$$

To establish that $\phi_2 \in \mathscr{T}$, choose

$$\begin{aligned} \kappa_{\mathcal{S}}(X) = 0, \qquad \kappa_{\mathcal{T}}(X) &= \frac{\widetilde{m}(X) - \mu}{\pi_{\mathcal{T}}}, \\ \zeta_{\mathcal{S}}(X,Y) &= \frac{q(X)a^*(X)}{\pi_{\mathcal{S}}}\{Y - m(X)\}, \\ \zeta_{\mathcal{T}}(X,Y,\widetilde{Y}) &= \frac{\widetilde{Y} - \widetilde{m}(X) - a^*(X)\{Y - m(X)\}}{\alpha\pi_{\mathcal{T}}}. \end{aligned}$$

These components are square-integrable and have the required mean-zero properties. Their conditional mean derivatives agree because

$$\frac{q(X)a^*(X)\sigma^2_{\mathcal{S},Y}(X)}{\pi_{\mathcal{S}}} = \frac{\gamma_{\mathcal{T}}(X) - a^*(X)\sigma^2_{\mathcal{T},Y}(X)}{\alpha\pi_{\mathcal{T}}} \quad P_{\mathcal{T},X}\text{-almost surely.}$$

Thus (16) holds and $\phi_2 \in \mathscr{T}$. Together with $\phi_2 = \phi_0 + h_{a^*}$ and $h_{a^*} \in \mathscr{T}^\perp$, this proves that ϕ_2 is the EIF of μ.

Since the EIF has mean zero in each sample, the semiparametric efficiency bound is

$$\begin{aligned} V_N &= \pi_{\mathcal{S}}\mathbb{E}_{\mathcal{S}}\{\phi_2(O)^2\} + \pi_{\mathcal{T}}\mathbb{E}_{\mathcal{T}}\{\phi_2(O)^2\} \\ &= \frac{\mathrm{Var}_{\mathcal{T}}\{\widetilde{m}(X)\}}{\pi_{\mathcal{T}}} + \mathbb{E}_{\mathcal{S}}\left[\frac{p_{\mathcal{T}}(X)^2a^*(X)^2\sigma^2_{\mathcal{S},Y}(X)}{\pi_{\mathcal{S}}p_{\mathcal{S}}(X)^2}\right] \\ &\quad + \frac{1}{\alpha\pi_{\mathcal{T}}}\mathbb{E}_{\mathcal{T}}\left[\sigma^2_{\mathcal{T},\widetilde{Y}}(X) - 2a^*(X)\gamma_{\mathcal{T}}(X) + a^*(X)^2\sigma^2_{\mathcal{T},Y}(X)\right] \\ &= \frac{\mathrm{Var}_{\mathcal{T}}\{\widetilde{m}(X)\}}{\pi_{\mathcal{T}}} + \frac{1}{\alpha\pi_{\mathcal{T}}}\mathbb{E}_{\mathcal{T}}\left[\sigma^2_{\mathcal{T},\widetilde{Y}}(X) - a^*(X)\gamma_{\mathcal{T}}(X)\right]. \end{aligned}$$

The last equality follows from

$$a^*(x)\left\{\sigma^2_{\mathcal{T},Y}(x) + \frac{\alpha\pi_{\mathcal{T}}p_{\mathcal{T}}(x)\sigma^2_{\mathcal{S},Y}(x)}{\pi_{\mathcal{S}}p_{\mathcal{S}}(x)}\right\} = \gamma_{\mathcal{T}}(x).$$

The corresponding first-order variance bound is V_N/N. If the constant α is unknown, its additional score direction is

$$s_\alpha(O) = \mathbb{I}(G = \mathcal{T})\frac{R-\alpha}{\alpha(1-\alpha)}.$$

Random selection for evaluation and conditional centering give $\pi_\mathcal{T}\mathbb{E}_\mathcal{T}\{\phi_1(O)s_\alpha(O)\} = \pi_\mathcal{T}\mathbb{E}_\mathcal{T}\{\phi_2(O)s_\alpha(O)\} = 0$. Since μ does not depend on α, both EIFs and their efficiency bounds remain unchanged when α is unknown.

D.3 Proof of Theorem 2

Write $n_{g,\ell} = |\mathcal{I}_{g,\ell}|$ and $N_\ell = N/L$, so $n_{g,\ell} = \pi_g N_\ell$ and $\widehat{\mu} = \sum_{\ell=1}^{L}(N_\ell/N)\widehat{\mu}_\ell$. Index the source records as $O_1, \ldots, O_{n_\mathcal{S}}$ and the target records as $O_{n_\mathcal{S}+1}, \ldots, O_N$.

We first show that

$$\sqrt{N_\ell}(\widehat{\mu}_\ell - \mu) = \frac{1}{\sqrt{N_\ell}}\left\{\sum_{i\in\mathcal{I}_{\mathcal{S},\ell}}\phi_2(O_i) + \sum_{j\in\mathcal{I}_{\mathcal{T},\ell}}\phi_2(O_{n_\mathcal{S}+j})\right\} + o_P(1). \tag{17}$$

Since L is fixed, this gives

$$\begin{aligned}\sqrt{N}(\widehat{\mu} - \mu) &= \sum_{\ell=1}^{L}\sqrt{\frac{N_\ell}{N}}\sqrt{N_\ell}(\widehat{\mu}_\ell - \mu)\\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi_2(O_i) + o_P(1).\end{aligned} \tag{18}$$

Fix ℓ, holding the fold partition and fitting randomization fixed. The full-sample weights make $\widehat{a}_\ell$ depend on evaluation covariates. To retain conditional centering of the evaluation residuals, we condition on all covariates as well as the training data $\mathcal{D}_{-\ell}$. Under this conditioning, all fitted quantities are fixed and the held-out outcomes and evaluation indicators remain independent across observations, with their conditional laws given X. Write $w^* = qa^*$ and suppress function arguments when unambiguous. Subtracting the oracle contributions gives

$$\begin{aligned}&\widehat{\mu}_\ell - \mathbb{P}_{\mathcal{T},\ell}\left[\widetilde{m} + \frac{R}{\alpha}\{\widetilde{Y} - \widetilde{m} - a^*(Y-m)\}\right] - \mathbb{P}_{\mathcal{S},\ell}[w^*(Y-m)]\\ &= \mathbb{P}_{\mathcal{T},\ell}\Delta_{\mathcal{T},\ell} + \mathbb{P}_{\mathcal{S},\ell}\Delta_{\mathcal{S},\ell},\end{aligned}$$

where

$$\begin{aligned}\Delta_{\mathcal{T},\ell} = &\left(1 - \frac{R}{\alpha}\right)(\widehat{\widetilde{m}}_\ell - \widetilde{m}) - \frac{R}{\alpha}(\widehat{a}_\ell - a^*)(Y-m) + \frac{R}{\alpha}\widehat{a}_\ell(\widehat{m}_\ell - m)\\ &+ R\left(\frac{1}{\widehat{\alpha}_\ell} - \frac{1}{\alpha}\right)\{\widetilde{Y} - \widehat{\widetilde{m}}_\ell - \widehat{a}_\ell(Y - \widehat{m}_\ell)\},\end{aligned} \tag{19}$$

and

$$\Delta_{\mathcal{S},\ell} = (\widehat{w}_\ell - w^*)(Y-m) - \widehat{w}_\ell(\widehat{m}_\ell - m). \tag{20}$$

The source expression is evaluated only at the observed source covariates. Define

$$\begin{aligned}\mathcal{B}_\ell &= \mathbb{E}[\mathbb{P}_{\mathcal{T},\ell}\Delta_{\mathcal{T},\ell} + \mathbb{P}_{\mathcal{S},\ell}\Delta_{\mathcal{S},\ell} \mid \mathcal{D}_{-\ell}, X_1, \ldots, X_N],\\ Z_\ell &= \sqrt{N_\ell}\{\mathbb{P}_{\mathcal{T},\ell}\Delta_{\mathcal{T},\ell} + \mathbb{P}_{\mathcal{S},\ell}\Delta_{\mathcal{S},\ell} - \mathcal{B}_\ell\}.\end{aligned}$$

It is enough to show that $\sqrt{N_\ell}\mathcal{B}_\ell = o_P(1)$ and $Z_\ell = o_P(1)$. Independent Bernoulli evaluation selection gives $\widehat{\alpha}_\ell - \alpha = O_P(N^{-1/2})$. By Assumptions 2–3, the last term in (19) contributes $o_P(N^{-1/2})$ to $\mathcal{B}_\ell$ and $O_P(N^{-1})$ to the conditional variance bound for Z_ℓ.

Since $\mathbb{E}_g[Y - m(X) \mid X] = 0$ for $g \in \{\mathcal{S}, \mathcal{T}\}$ and $\mathbb{E}_\mathcal{T}[R \mid X, Y, \widetilde{Y}] = \alpha$, taking expectations conditional on $\mathcal{D}_{-\ell}$ and all covariates gives

$$\begin{aligned}\mathcal{B}_\ell = &\mathbb{P}_{\mathcal{T},\ell}[\widehat{a}_\ell(\widehat{m}_\ell - m)] - \mathbb{P}_{\mathcal{S},\ell}[q\widehat{a}_\ell(\widehat{m}_\ell - m)]\\ &- \mathbb{P}_{\mathcal{S},\ell}[(\widehat{q} - q)\widehat{a}_\ell(\widehat{m}_\ell - m)] + o_P(N^{-1/2}).\end{aligned}$$

By the reproducing property, $\|\widehat{a}_\ell\|_\infty \le \sqrt{K_0}\|\widehat{a}_\ell\|_{\mathcal{H}} = O_P(1)$. To control the density-ratio error term, separate its population mean from its empirical fluctuation. Cauchy–Schwarz and Assumption 4 give

$$|\mathbb{E}_{\mathcal{S}}[(\widehat{q}-q)\widehat{a}_\ell(\widehat{m}_\ell - m)]| \le \|\widehat{a}_\ell\|_\infty \|\widehat{q}-q\|_{2,\mathcal{S}}\|\widehat{m}_\ell - m\|_{2,\mathcal{S}} = o_P(N^{-1/2}).$$

For every fixed $C < \infty$, the class $\{(f-q)a : f \in \mathcal{Q},\ \|a\|_{\mathcal{H}} \le C\}$ is uniformly bounded and $P_{\mathcal{S}}$-Donsker, since bounded RKHS balls are Donsker and bounded products preserve this property. Conditional on $\mathcal{D}_{-\ell}$, the multiplier $\widehat{m}_\ell - m$ is fixed, bounded in probability, and converges to zero in $L_2(P_{\mathcal{S}})$. A finite $L_2(P_{\mathcal{S}})$ approximation, stochastic equicontinuity, and the multiplier contraction inequality therefore yield

$$\sqrt{n_{\mathcal{S},\ell}} \sup_{\substack{f\in\mathcal{Q}\\ \|a\|_{\mathcal{H}}\le C}} |(\mathbb{P}_{\mathcal{S},\ell} - \mathbb{E}_{\mathcal{S}})[(f-q)a(\widehat{m}_\ell - m)]| = o_P(1).$$

Together with $\|\widehat{a}_\ell\|_{\mathcal{H}} = O_P(1)$, these bounds give $\mathbb{P}_{\mathcal{S},\ell}[(\widehat{q}-q)\widehat{a}_\ell(\widehat{m}_\ell - m)] = o_P(N^{-1/2})$. For the remaining covariate discrepancy, the same reproducing property gives the pathwise bound

$$\begin{aligned}&|\mathbb{P}_{\mathcal{T},\ell}[\widehat{a}_\ell(\widehat{m}_\ell - m)] - \mathbb{P}_{\mathcal{S},\ell}[q\widehat{a}_\ell(\widehat{m}_\ell - m)]|\\ \le&\|\widehat{a}_\ell\|_{\mathcal{H}}\,\|\mathbb{P}_{\mathcal{T},\ell}[(\widehat{m}_\ell - m)\mathsf{K}(X,\cdot)] - \mathbb{P}_{\mathcal{S},\ell}[q(\widehat{m}_\ell - m)\mathsf{K}(X,\cdot)]\|_{\mathcal{H}}\,.\end{aligned}$$

Conditional on $\mathcal{D}_{-\ell}$, the regression $\widehat{m}_\ell$ is fixed and the held-out samples are independent. The identity $\mathbb{E}_{\mathcal{S}}[q(X)f(X)] = \mathbb{E}_{\mathcal{T}}[f(X)]$ implies that the difference of kernel averages below has conditional mean zero. Hence, using $\mathsf{K}(X,X) \le K_0$,

$$\begin{aligned}&\mathbb{E}\Big[\|\mathbb{P}_{\mathcal{T},\ell}[(\widehat{m}_\ell - m)\mathsf{K}(X,\cdot)] - \mathbb{P}_{\mathcal{S},\ell}[q(\widehat{m}_\ell - m)\mathsf{K}(X,\cdot)]\|_{\mathcal{H}}^2 \,\Big|\, \mathcal{D}_{-\ell}\Big]\\ \le&\frac{K_0}{n_{\mathcal{T},\ell}}\|\widehat{m}_\ell - m\|_{2,\mathcal{T}}^2 + \frac{K_0}{n_{\mathcal{S},\ell}}\|q(\widehat{m}_\ell - m)\|_{2,\mathcal{S}}^2 = o_P(N^{-1}),\end{aligned}$$

where the last equality follows from boundedness of q and Assumption 3. By conditional Markov's inequality, the norm is $o_P(N^{-1/2})$. Combining this with $\|\widehat{a}_\ell\|_{\mathcal{H}} = O_P(1)$ and the preceding bound on the density-ratio error term gives $\mathcal{B}_\ell = o_P(N^{-1/2})$.

We next control Z_ℓ. By definition, $\mathbb{E}[Z_\ell \mid \mathcal{D}_{-\ell}, X_1, \dots, X_N] = 0$. Conditional independence of the held-out observations, boundedness of $Y - m$, and (19)–(20) imply

$$\begin{aligned}&\mathrm{Var}(Z_\ell \mid \mathcal{D}_{-\ell}, X_1, \dots, X_N)\\ \lesssim&\mathbb{P}_{\mathcal{T},\ell}[(\widehat{\widetilde{m}}_\ell - \widetilde{m})^2] + \mathbb{P}_{\mathcal{T},\ell}[(\widehat{a}_\ell - a^*)^2] + \mathbb{P}_{\mathcal{S},\ell}[(\widehat{w}_\ell - w^*)^2]\\ &+ \|\widehat{a}_\ell\|_\infty^2\mathbb{P}_{\mathcal{T},\ell}[(\widehat{m}_\ell - m)^2] + O_P(N^{-1}) = o_P(1).\end{aligned} \tag{21}$$

The empirical regression errors are $o_P(1)$ by Assumption 3 and conditional Markov's inequality given $\mathcal{D}_{-\ell}$. The empirical coefficient errors are $o_P(1)$ by the same assumption, and $\|\widehat{a}_\ell\|_\infty = O_P(1)$ as shown above. Conditional Chebyshev's inequality therefore gives $Z_\ell = o_P(1)$. Together with $\sqrt{N_\ell}\mathcal{B}_\ell = o_P(1)$, this establishes (17) and (18).

For variance estimation, we also control the empirical squared score errors. In particular,

$$\mathbb{P}_{\mathcal{S},\ell}[\widehat{w}_\ell^2(\widehat{m}_\ell - m)^2] \le \left\{\max_{i\le n_{\mathcal{S}}} \widehat{q}(X_i^{\mathcal{S}})\right\}^2 \|\widehat{a}_\ell\|_\infty^2\mathbb{P}_{\mathcal{S},\ell}[(\widehat{m}_\ell - m)^2] = o_P(1).$$

The last term in (19) also has empirical second moment $O_P(N^{-1})$. Together with (19)–(20) and bounded R/α and $Y - m$, this yields

$$\max_{\ell\le L}\{\mathbb{P}_{\mathcal{T},\ell}(\Delta_{\mathcal{T},\ell}^2) + \mathbb{P}_{\mathcal{S},\ell}(\Delta_{\mathcal{S},\ell}^2)\} = o_P(1). \tag{22}$$

The overall empirical fraction $\widehat{\alpha} = n_{\mathcal{T}}^{-1}\sum_{j=1}^{n_{\mathcal{T}}} R_j$ can also be used in the held-out scores without changing these limits. Indeed, $\widehat{\alpha} - \widehat{\alpha}_\ell = O_P(N^{-1/2})$, and conditional centering and regression consistency give

$$\left(\frac{1}{\widehat{\alpha}} - \frac{1}{\widehat{\alpha}_\ell}\right)\mathbb{P}_{\mathcal{T},\ell}\Big[R\{\widetilde{Y} - \widehat{\widetilde{m}}_\ell - \widehat{a}_\ell(Y - \widehat{m}_\ell)\}\Big] = O_P(N^{-1/2})o_P(1) = o_P(N^{-1/2}).$$

The corresponding empirical score difference is also $o_P(1)$ in squared mean, so the variance consistency argument below applies to either normalization.

To establish the limiting normal law in (12), write $V_N = \mathrm{Var}\{\phi_2(O)\}$ and first show $V_N = V + o(1)$. Write $\pi = (\pi_{\mathcal{S}}, \pi_{\mathcal{T}})$ and $\pi^0 = (\pi^0_{\mathcal{S}}, \pi^0_{\mathcal{T}})$, and make the design dependence explicit in $a^*(x;\pi)$, $w^*(x;\pi)$, and $\phi_2(O;\pi)$. In this part of the proof, let $a^0(x) = a^*(x;\pi^0)$ and $w^0(x) = w^*(x;\pi^0)$. The coefficient formulas in Proposition 1 give

$$a^0(x)\left\{\sigma^2_{\mathcal{T},Y}(x) + \frac{\alpha\pi^0_{\mathcal{T}} p_{\mathcal{T}}(x)}{\pi^0_{\mathcal{S}} p_{\mathcal{S}}(x)}\sigma^2_{\mathcal{S},Y}(x)\right\} = \gamma_{\mathcal{T}}(x).$$

Multiplying by $a^0(x)$, integrating over the target covariate law, and using $w^0 = (p_{\mathcal{T}}/p_{\mathcal{S}})a^0$, we obtain

$$\begin{aligned}&\frac{1}{\pi^0_{\mathcal{S}}}\mathbb{E}_{\mathcal{S}}[(w^0)^2\sigma^2_{\mathcal{S},Y}] + \frac{1}{\alpha\pi^0_{\mathcal{T}}}\mathbb{E}_{\mathcal{T}}[(a^0)^2\sigma^2_{\mathcal{T},Y}]\\ &= \frac{1}{\alpha\pi^0_{\mathcal{T}}}\mathbb{E}_{\mathcal{T}}[a^0\gamma_{\mathcal{T}}] \le \frac{1}{\alpha\pi^0_{\mathcal{T}}}\mathbb{E}_{\mathcal{T}}[\sigma^2_{\mathcal{T},\widetilde{Y}}] < \infty.\end{aligned}$$

Here conditional Cauchy–Schwarz gives $\gamma^2_{\mathcal{T}} \le \sigma^2_{\mathcal{T},\widetilde{Y}}\sigma^2_{\mathcal{T},Y}$ and $0 \le a^0\gamma_{\mathcal{T}} \le \sigma^2_{\mathcal{T},\widetilde{Y}}$, and finiteness follows from Assumption 2. Both terms on the left are nonnegative and therefore finite.

Since $\pi_g/\pi^0_g = 1 + o(1)$, for almost every fixed $x \in \mathcal{X}$,

$$\begin{aligned}a^*(x;\pi) &= a^0(x) + o(1), \qquad w^*(x;\pi) = w^0(x) + o(1),\\ |a^*(x;\pi)| &\lesssim |a^0(x)|, \qquad |w^*(x;\pi)| \lesssim |w^0(x)|.\end{aligned}$$

The coefficient denominators are sums of nonnegative terms, so the bounds hold for all sufficiently large N, with implicit constants independent of N and x. In particular, $\phi_2(O;\pi) \to \phi_2(O;\pi^0)$ almost surely.

For $G = \mathcal{S}$, (2) yields

$$\begin{aligned}&\{\phi_2(O;\pi) - \phi_2(O;\pi^0)\}^2\\ &= \left\{\frac{w^*(X;\pi)}{\pi_{\mathcal{S}}} - \frac{w^*(X;\pi^0)}{\pi^0_{\mathcal{S}}}\right\}^2 (Y-m)^2\\ &\lesssim (w^0)^2(Y-m)^2.\end{aligned}$$

For $G = \mathcal{T}$, Cauchy–Schwarz gives

$$\begin{aligned}&\{\phi_2(O;\pi) - \phi_2(O;\pi^0)\}^2\\ &\lesssim (\widetilde{m}-\mu)^2 + R(\widetilde{Y}-\widetilde{m})^2 + R(a^0)^2(Y-m)^2.\end{aligned}$$

These pointwise bounds have constants independent of N. Their right-hand sides do not depend on N and are integrable, since independent selection for evaluation gives

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}[(w^0)^2(Y-m)^2] + \mathbb{E}_{\mathcal{T}}[(\widetilde{m}-\mu)^2 + R(\widetilde{Y}-\widetilde{m})^2 + R(a^0)^2(Y-m)^2]\\ &= \mathbb{E}_{\mathcal{S}}[(w^0)^2\sigma^2_{\mathcal{S},Y}] + \mathrm{Var}_{\mathcal{T}}(\widetilde{m}) + \alpha\mathbb{E}_{\mathcal{T}}[\sigma^2_{\mathcal{T},\widetilde{Y}}] + \alpha\mathbb{E}_{\mathcal{T}}[(a^0)^2\sigma^2_{\mathcal{T},Y}] < \infty.\end{aligned}$$

Combining these integrable bounds with pointwise convergence, the dominated convergence theorem gives, for $g \in \{\mathcal{S}, \mathcal{T}\}$,

$$\begin{aligned}&\lim_{N\to\infty}\mathbb{E}_g\left[\{\phi_2(O;\pi) - \phi_2(O;\pi^0)\}^2\right]\\ &= \mathbb{E}_g\left[\lim_{N\to\infty}\{\phi_2(O;\pi) - \phi_2(O;\pi^0)\}^2\right] = 0.\end{aligned}$$

The same bounds give $\|\phi_2(\cdot;\pi)\|_{2,g} + \|\phi_2(\cdot;\pi^0)\|_{2,g} = O(1)$. By Cauchy–Schwarz,

$$\begin{aligned}&\left|\mathbb{E}_g[\{\phi_2(O;\pi)\}^2] - \mathbb{E}_g[\{\phi_2(O;\pi^0)\}^2]\right|\\ &\le \|\phi_2(\cdot;\pi) - \phi_2(\cdot;\pi^0)\|_{2,g}\left\{\|\phi_2(\cdot;\pi)\|_{2,g} + \|\phi_2(\cdot;\pi^0)\|_{2,g}\right\}\\ &= o(1).\end{aligned}$$

Consequently, the efficiency bound in Proposition 1 satisfies

$$\begin{aligned} V_N &= \sum_{g\in\{\mathcal{S},\mathcal{T}\}} \pi_g \mathbb{E}_g[\{\phi_2(O;\pi)\}^2] \\ &= \sum_{g\in\{\mathcal{S},\mathcal{T}\}} \pi_g^0 \mathbb{E}_g[\{\phi_2(O;\pi^0)\}^2] + o(1) \\ &= V + o(1). \end{aligned}$$

Furthermore,

$$\mathbb{E}_g[\phi_2(O;\pi)] = \mathbb{E}_g[\phi_2(O;\pi^0)] = 0, \qquad g \in \{\mathcal{S},\mathcal{T}\}.$$

Hence,

$$\begin{aligned} &\mathbb{E}\left[\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[\phi_2(O_i;\pi) - \phi_2(O_i;\pi^0)]\right\}^2\right] \\ &= \frac{1}{N}\sum_{g\in\{\mathcal{S},\mathcal{T}\}} n_g \mathbb{E}_g[\{\phi_2(O;\pi) - \phi_2(O;\pi^0)\}^2] \\ &= \sum_{g\in\{\mathcal{S},\mathcal{T}\}} \pi_g \mathbb{E}_g[\{\phi_2(O;\pi) - \phi_2(O;\pi^0)\}^2] = o(1). \end{aligned}$$

By the central limit theorem within each group and independence between the groups,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi_2(O_i;\pi^0) = \sum_{g\in\{\mathcal{S},\mathcal{T}\}}\sqrt{\pi_g}\left\{\frac{1}{\sqrt{n_g}}\sum_{i:G_i=g}\phi_2(O_i;\pi^0)\right\} \xrightarrow{d} \mathcal{N}(0,V).$$

Therefore, by (18), we have

$$\sqrt{N}(\widehat{\mu}-\mu) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi_2(O_i) + o_P(1) \xrightarrow{d} \mathcal{N}(0,V).$$

To prove consistency of (11), define $\widehat{\phi}_{2,\ell}$ on the evaluation observations by the EIF in (2) with $\widetilde{m}$, m, a^*, $(p_{\mathcal{T}}/p_{\mathcal{S}})a^*$, α, and μ replaced by $\widehat{\widetilde{m}}_\ell$, $\widehat{m}_\ell$, $\widehat{a}_\ell$, $\widehat{w}_\ell$, $\widehat{\alpha}_\ell$, and $\widehat{\mu}$, respectively. Then

$$\widehat{\phi}_{2,\ell}(O) - \phi_2(O;\pi) = \frac{\mathbb{I}(G=\mathcal{T})}{\pi_{\mathcal{T}}}\{\Delta_{\mathcal{T},\ell}(O^{\mathcal{T}},R) - (\widehat{\mu}-\mu)\} + \frac{\mathbb{I}(G=\mathcal{S})}{\pi_{\mathcal{S}}}\Delta_{\mathcal{S},\ell}(O^{\mathcal{S}}).$$

For each $g \in \{\mathcal{S},\mathcal{T}\}$ and $\ell \le L$, (22) already gives $\mathbb{P}_{g,\ell}(\Delta_{g,\ell}^2) = o_P(1)$. The preceding L_2 convergence and Markov's inequality also give

$$\mathbb{P}_{g,\ell}\big[\{\phi_2(O;\pi) - \phi_2(O;\pi^0)\}^2\big] = o_P(1).$$

Together with $\widehat{\mu}-\mu = o_P(1)$ and $\pi_g = \pi_g^0 + o(1)$, these bounds imply

$$\begin{aligned} &\mathbb{P}_{g,\ell}\Big[\{\widehat{\phi}_{2,\ell}(O) - \phi_2(O;\pi^0)\}^2\Big] \\ &\lesssim \mathbb{P}_{g,\ell}(\Delta_{g,\ell}^2) + (\widehat{\mu}-\mu)^2 + \mathbb{P}_{g,\ell}\big[\{\phi_2(O;\pi) - \phi_2(O;\pi^0)\}^2\big] = o_P(1). \end{aligned}$$

By square integrability and the law of large numbers,

$$\mathbb{P}_{g,\ell}\big[\{\phi_2(O;\pi^0)\}^2\big] = \mathbb{E}_g\big[\{\phi_2(O;\pi^0)\}^2\big] + o_P(1) = O_P(1).$$

Cauchy–Schwarz and these bounds imply

$$\Big|\mathbb{P}_{g,\ell}\Big[\{\widehat{\phi}_{2,\ell}(O)\}^2 - \{\phi_2(O;\pi^0)\}^2\Big]\Big| = o_P(1).$$

Therefore,

$$\begin{aligned} \widehat{V} &= \frac{1}{L}\sum_{\ell=1}^{L}\sum_{g\in\{\mathcal{S},\mathcal{T}\}}\pi_g\mathbb{P}_{g,\ell}\Big[\{\widehat{\phi}_{2,\ell}(O)\}^2\Big] \\ &= \sum_{g\in\{\mathcal{S},\mathcal{T}\}}\pi_g^0\mathbb{E}_g\big[\{\phi_2(O;\pi^0)\}^2\big] + o_P(1) = V + o_P(1). \end{aligned}$$

Along any regular parametric submodel, contiguity preserves the remainder in (18), and Le Cam's third lemma shifts the oracle sum by the pathwise derivative of μ. Centering at the corresponding local parameter therefore gives the same limiting normal law. The estimator is regular, and Proposition 1 identifies its influence function as the EIF.

E Details of Experiments

E.1 Evaluation of Semi-synthetic Data

We conduct a HealthBench-calibrated semi-synthetic study to assess the finite-sample performance of the proposed estimators under conditional-mean invariance $\mathbb{E}_{\mathcal{S}}[Y \mid X] = \mathbb{E}_{\mathcal{T}}[Y \mid X]$. We retain the observed covariates and generate evaluator scores whose marginal distributions resemble those in the real data. The source and target domains share the same conditional mean of the benchmark-evaluator score Y, but have different conditional variances, allowing us to examine performance without imposing equality of the conditional distributions. We consider both the same-evaluator setting ($Y = \widetilde{Y}$) and the different-evaluator setting ($Y \neq \widetilde{Y}$). In the latter, we vary the marginal correlation between Y and $\widetilde{Y}$ over 0.4, 0.6, and 0.8. For reference, the GPT-4.1 and Gemini Flash-Lite scores on the 5,000 responses had a marginal Pearson correlation of 0.793 and a residual correlation of 0.756, with residuals computed using five-fold out-of-fold predictions on the original score scale. Thus, the correlation setting of 0.8 approximates the association observed in the real data, whereas 0.4 and 0.6 represent weaker associations. Across target-evaluation budgets of 200, 300, and 500, we compare estimation accuracy using 500 repeated source–target partitions of each fixed simulated dataset.

E.1.1 Simulation Setting

We constructed a semi-synthetic experiment using 5,000 HealthBench examples, with Gemini 3.1 Pro as the response model. We used the same 500 source–target splits as in Case 2, each containing 1,439 source and 3,561 target examples. We retained the fixed PC50 prompt–rubric representations as covariates $X_i \in \mathbb{R}^{80}$, comprising 58 prompt principal components and 22 rubric principal components. Scores assigned to the model's responses by GPT-4.1 and Gemini Flash-Lite were used to calibrate the benchmark-evaluator score Y and, in the different-evaluator settings, the target score $\widetilde{Y}$, respectively. Let $Y^{\mathcal{T}}$ and $\widetilde{Y}^{\mathcal{T}}$ denote the simulated benchmark and *target evaluator* scores, respectively, for a target-domain example.

We transformed each evaluator's scores to a normal-score scale, fitted ridge regressions on the covariates, and added Gaussian noise with scales estimated from out-of-fold residuals. The generated values were then mapped back to the original score scale using quantile mappings calibrated to each evaluator's empirical score distribution. In the different-evaluator settings, we adjusted the noise correlation so that the empirical correlation between $\widetilde{Y}^{\mathcal{T}}$ and $Y^{\mathcal{T}}$ across all 5,000 examples was 0.4, 0.6, or 0.8, within 10^{-4}.

To impose conditional-mean invariance, let $m(x) = \mathbb{E}_{\mathcal{T}}[Y \mid X = x]$ under the fitted generator. We construct the source benchmark-evaluator score $Y^{\mathcal{S}}$ as

$$Y^{\mathcal{S}} = m(X) + \frac{1}{2}\{Y^{\mathcal{T}} - m(X)\}.$$

Thus, $\mathbb{E}_{\mathcal{S}}[Y \mid X] = \mathbb{E}_{\mathcal{T}}[Y \mid X] = m(X)$, while the source conditional variance is one quarter of the target conditional variance. In the same-evaluator setting, both evaluators use $Y^{\mathcal{S}}$ in the source domain and $Y^{\mathcal{T}}$ in the target domain.

Each setting's simulated scores were generated once and held fixed across 500 source–target partitions, each containing 1,439 source and 3,561 target examples. Within each partition, target tasks were selected for evaluation uniformly without replacement using a shared random ordering, yielding nested budgets of 200, 300, and 500 target evaluations for all methods. The evaluation truth for each split was the average of $\widetilde{m}(X_i)$ over its target examples, where $\widetilde{m}(x) = \mathbb{E}_{\mathcal{T}}[\widetilde{Y} \mid X = x]$. These conditional means were computed under the fitted generator and withheld from the estimators.

E.1.2 Additional Simulation Results

Tables 5–8 report estimation accuracy over 500 repeated splits of each fixed simulated dataset, using the split-specific truth defined in Section E.1.1. We report bias, mean absolute error (MAE), and root mean squared error (RMSE), each multiplied by 100, for 200, 300, and 500 target scores. Green and blue identify the best and second-best displayed values, respectively; bias is ranked by absolute value, and ties share a color.

Across settings, biases are small relative to RMSEs. CAFE exhibits slightly larger bias than some baselines in several different-evaluator settings, but the increase is small relative to its reduction in variability, yielding lower RMSE as the marginal correlation between $\widetilde{Y}$ and Y strengthens.

Table 5: Same-evaluator semi-synthetic setting. Columns give target-evaluation counts; parentheses in methods identify samples supplying observed scores, with $\mathcal{L} = \{j, R_j = 1\}$ indexing evaluated target tasks. Green/blue denote ranks 1/2; bias is ranked by absolute value.

Method	Bias ($\times 10^2$)			MAE ($\times 10^2$) $\downarrow$			RMSE ($\times 10^2$) $\downarrow$		
	200	300	500	200	300	500	200	300	500
Target-only sample mean ($\mathcal{L}$)	-0.149	-0.088	-0.059	1.78	1.43	1.04	2.21	1.77	1.32
PPI++ ($\mathcal{S}, \mathcal{L}$)	-0.204	-0.128	-0.094	1.63	1.31	0.96	2.02	1.63	1.22
Target-only AIPW ($\mathcal{L}$)	-0.143	-0.106	-0.077	1.78	1.41	1.01	2.22	1.76	1.29
Pooled-label AIPW ($\mathcal{S}, \mathcal{L}$)	-0.124	-0.092	-0.081	1.60	1.31	0.97	2.01	1.64	1.23
Pooled-label DR ($\mathcal{S}, \mathcal{L}$)	-0.080	-0.085	-0.120	0.56	0.56	0.56	0.70	0.71	0.71
CAFE ($\mathcal{S}, \mathcal{L}$)	0.029	0.035	0.025	0.44	0.42	0.38	0.54	0.53	0.49

Table 6: Different-evaluator semi-synthetic setting with calibrated marginal correlation 0.4 between $\widetilde{Y}$ and Y. Columns give target-evaluation counts; parentheses in methods identify samples supplying observed scores, with $\mathcal{L} = \{j, R_j = 1\}$ indexing evaluated target tasks. Green/blue denote ranks 1/2; bias is ranked by absolute value.

Method	Bias ($\times 10^2$)			MAE ($\times 10^2$) $\downarrow$			RMSE ($\times 10^2$) $\downarrow$		
	200	300	500	200	300	500	200	300	500
Target-only sample mean ($\mathcal{L}$)	0.655	0.644	0.669	1.77	1.45	1.15	2.20	1.79	1.44
PPI++ ($\mathcal{S}, \mathcal{L}$)	0.607	0.609	0.641	1.69	1.38	1.09	2.09	1.70	1.36
RePPI ($\mathcal{S}, \mathcal{L}$)	0.621	0.601	0.638	1.75	1.41	1.12	2.19	1.74	1.39
Target-only AIPW ($\mathcal{L}$)	0.670	0.641	0.643	1.78	1.43	1.11	2.20	1.77	1.39
CAFE ($\mathcal{S}, \mathcal{L}$)	0.710	0.673	0.671	1.68	1.37	1.07	2.10	1.69	1.34

Table 7: Different-evaluator semi-synthetic setting with calibrated marginal correlation 0.6 between $\widetilde{Y}$ and Y. Columns give target-evaluation counts; parentheses in methods identify samples supplying observed scores, with $\mathcal{L} = \{j, R_j = 1\}$ indexing evaluated target tasks. Green/blue denote ranks 1/2; bias is ranked by absolute value.

Method	Bias ($\times 10^2$)			MAE ($\times 10^2$) $\downarrow$			RMSE ($\times 10^2$) $\downarrow$		
	200	300	500	200	300	500	200	300	500
Target-only sample mean ($\mathcal{L}$)	0.503	0.507	0.538	1.81	1.46	1.13	2.23	1.82	1.42
PPI++ ($\mathcal{S}, \mathcal{L}$)	0.456	0.471	0.509	1.73	1.40	1.06	2.12	1.72	1.34
RePPI ($\mathcal{S}, \mathcal{L}$)	0.464	0.470	0.503	1.80	1.44	1.10	2.23	1.77	1.38
Target-only AIPW ($\mathcal{L}$)	0.518	0.504	0.508	1.82	1.46	1.09	2.25	1.80	1.38
CAFE ($\mathcal{S}, \mathcal{L}$)	0.597	0.563	0.560	1.55	1.26	0.97	1.92	1.57	1.22

Table 8: Different-evaluator semi-synthetic setting with calibrated marginal correlation 0.8 between $\widetilde{Y}$ and Y. Columns give target-evaluation counts; parentheses in methods identify samples supplying observed scores, with $\mathcal{L} = \{j, R_j = 1\}$ indexing evaluated target tasks. Green/blue denote ranks 1/2; bias is ranked by absolute value.

Method	Bias ($\times 10^2$)			MAE ($\times 10^2$) ↓			RMSE ($\times 10^2$) ↓		
	200	300	500	200	300	500	200	300	500
Target-only sample mean ($\mathcal{L}$)	0.284	0.310	0.343	1.83	1.47	1.11	2.26	1.84	1.42
PPI++ ($\mathcal{S}, \mathcal{L}$)	0.236	0.273	0.312	1.74	1.40	1.05	2.13	1.73	1.33
RePPI ($\mathcal{S}, \mathcal{L}$)	0.232	0.272	0.311	1.81	1.45	1.09	2.24	1.80	1.38
Target-only AIPW ($\mathcal{L}$)	0.300	0.308	0.312	1.85	1.46	1.10	2.28	1.82	1.39
CAFE ($\mathcal{S}, \mathcal{L}$)	0.423	0.397	0.390	1.28	1.04	0.81	1.59	1.29	1.02

E.2 The Same-Evaluator Setting

E.2.1 Implementation Details of CAFE

When the same evaluator is used in the source and target populations, so that $Y = \widetilde{Y}$, we fit a shared outcome regression using the pooled source and evaluated target observations. The regression is fitted with ridge regression. The density ratio $q = p_{\mathcal{T}}/p_{\mathcal{S}}$ is estimated by KMM using all source and target covariates, motivated by its comparatively stable performance in our task-distribution experiments. The source weights are normalized to mean one and held fixed across folds.

We implement CAFE using five-fold cross-fitting. Within each outer training fold, we choose $\lambda_a \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ by three-fold cross-validation, with target observations stratified by evaluation status. For each candidate value, the shared outcome regression and coefficient function are fitted on the inner training folds and evaluated using the held-out criterion corresponding to (13). We select the value with the lowest average validation criterion and refit the coefficient function using the resulting out-of-fold residuals from the full outer training sample.

We model $a(x)$ in an RKHS with Gaussian kernel

$$k(x, x') = \exp\left\{-\frac{\|x - x'\|_2^2}{2\sigma^2}\right\},$$

where σ is set to the median pairwise Euclidean distance among at most 2,000 pooled source and target observations in the corresponding training sample. The target-evaluation fraction is computed from the data used to fit each coefficient function, while the final held-out scores use the overall empirical fraction $\widehat{\alpha} = n_{\mathcal{L}}/n_{\mathcal{T}}$.

E.2.2 Implementation Details of Baselines

For comparison, we include target-only sample mean, source reweighting, PPI++, target-only AIPW, pooled-label AIPW, and pooled-label DR as baselines. Unless otherwise specified, regression models used by the baselines are fitted with ridge regression and density ratios are estimated with KMM.

Target-only sample mean. The target-only sample mean uses only observed target outcomes and requires no nuisance estimation:

$$\widehat{\mu}_{\mathrm{TO}} = \mathbb{P}_{\mathcal{L}} Y.$$

It therefore provides a benchmark for the information contained in the target scores alone.

Source reweighting. Source reweighting uses the observed source scores together with source and target covariates. We estimate the target-to-source density ratio $q(x)$ by KMM and transport the source outcome mean to the target distribution:

$$\widehat{\mu}_{\mathrm{SR}} = \frac{\mathbb{P}_{\mathcal{S}}[\widehat{q}(X)Y]}{\mathbb{P}_{\mathcal{S}}\widehat{q}(X)}.$$

PPI++. PPI++ uses the source sample from $\mathcal{S}$ to construct predictions and the evaluated target sample $\mathcal{L}$ to calibrate their contribution. We fit a source predictor $\widehat{f}_{\mathcal{S}}$ on (X, Y) from $\mathcal{S}$ and estimate

$$\widehat{\mu}_{\mathrm{PPI++}} = \mathbb{P}_{\mathcal{L}} Y + \widehat{\lambda} \left\{ \mathbb{P}_{\mathcal{U}} \widehat{f}_{\mathcal{S}}(X) - \mathbb{P}_{\mathcal{L}} \widehat{f}_{\mathcal{S}}(X) \right\},$$

where $\mathcal{L}$ denotes the examples with target scores and $\mathcal{U}$ denotes the examples without target scores. For each response model and source–target split, the source predictor is fitted once and its predictions are reused across target-evaluation budgets. The power coefficient $\widehat{\lambda}$ is re-estimated separately at each target-evaluation budget using the evaluated target outcomes and corresponding predictions to minimize the estimated asymptotic variance.

Target-only AIPW. Target-only AIPW uses only observations from $\mathcal{T}$. We implement five-fold cross-fitting. For each evaluation fold ℓ, the regression function $\widehat{f}_{\mathcal{T},\ell}$ is fitted using evaluated target observations in the complementary training folds and is then evaluated on the held-out target fold. The estimator is

$$\widehat{\mu}_{\mathrm{AIPW}} = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{P}_{\mathcal{T},\ell} \left[\widehat{f}_{\mathcal{T},\ell}(X) + \frac{R}{\widehat{\alpha}} \left\{ Y - \widehat{f}_{\mathcal{T},\ell}(X) \right\} \right],$$

where $L = 5$ and $\widehat{\alpha} = n_{\mathcal{L}}/n_{\mathcal{T}}$ is the overall empirical target-evaluation fraction.

Pooled-label AIPW. Pooled-label AIPW augments the target-only AIPW estimator by using source outcomes when learning the outcome regression. We again implement five-fold cross-fitting. For each evaluation fold ℓ, $\widehat{m}_{\ell}$ is fitted using the pooled source and evaluated-target observations, $\mathcal{S} \cup \mathcal{L}$, outside the corresponding evaluation fold, and is then evaluated on the held-out target observations. We estimate

$$\widehat{\mu}_{\mathrm{PAIPW}} = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{P}_{\mathcal{T},\ell} \left[\widehat{m}_{\ell}(X) + \frac{R}{\widehat{\alpha}} \left\{ Y - \widehat{m}_{\ell}(X) \right\} \right],$$

where $L = 5$ and $\widehat{\alpha} = n_{\mathcal{L}}/n_{\mathcal{T}}$. Relative to target-only AIPW, the source observations provide additional information for estimating the outcome regression, while the residual correction continues to use only observed target outcomes.

Pooled-label DR. Pooled-label DR uses both source and evaluated-target outcomes in the residual correction and accounts for the source–target covariate shift. We fit KMM weights $\widehat{v}_i$ directly on the covariates of the pooled evaluated sample, $\mathcal{S} \cup \mathcal{L}$, against all target covariates. These weights approximate the target density divided by the pooled evaluated-sample density, are normalized to mean one, and are held fixed across folds. With five-fold cross-fitting, the pooled outcome regression $\widehat{m}_{\ell}$ is fitted using source and evaluated-target outcomes outside evaluation fold ℓ. Let

$$\mathcal{A}_{\ell} = \mathcal{I}_{\mathcal{S},\ell} \cup (\mathcal{L} \cap \mathcal{I}_{\mathcal{T},\ell}),$$

where $\mathcal{A}_{\ell}$ contains the source and evaluated-target observations in evaluation fold ℓ. Its self-normalized estimator is

$$\widehat{\mu}_{\mathrm{pool}} = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{P}_{\mathcal{T},\ell} \widehat{m}_{\ell}(X) + \frac{\displaystyle\sum_{\ell=1}^{L} \sum_{i \in \mathcal{A}_{\ell}} \widehat{v}_i \left\{ Y_i - \widehat{m}_{\ell}(X_i) \right\}}{\displaystyle\sum_{\ell=1}^{L} \sum_{i \in \mathcal{A}_{\ell}} \widehat{v}_i}.$$

The first term averages the pooled regression predictions over the target distribution, while the second term uses density-ratio-weighted residuals from both source and evaluated-target observations. This estimator adapts the distribution-aligned fusion construction of Qiu et al. (2024) to the partially evaluated target setting.

E.2.3 Additional Results

Table 9: Same-evaluator setting in Case 1. Columns give target-evaluation counts; parentheses in methods identify samples supplying observed scores, with $\mathcal{L} = \{j, R_j = 1\}$ indexing evaluated target tasks. Green/blue denote ranks 1/2.

Method	Full rank recovery ↑			Kendall's τ ↑			MAE ($\times 10^2$) ↓			RMSE ($\times 10^2$) ↓		
	200	300	500	200	300	500	200	300	500	200	300	500
Target-only sample mean ($\mathcal{L}$)	0.38	0.42	0.52	0.83	0.85	0.88	1.59	1.32	0.99	1.82	1.50	1.12
Source reweighting ($\mathcal{S}$)	0.50	0.50	0.50	0.86	0.86	0.86	1.64	1.64	1.64	1.83	1.83	1.83
PPI++ ($\mathcal{S}, \mathcal{L}$)	0.40	0.42	0.50	0.83	0.84	0.88	1.53	1.23	0.94	1.76	1.42	1.08
Target-only AIPW ($\mathcal{L}$)	0.34	0.42	0.49	0.81	0.84	0.87	1.58	1.28	0.94	1.82	1.46	1.08
Pooled-label AIPW ($\mathcal{S}, \mathcal{L}$)	0.37	0.42	0.50	0.82	0.84	0.88	1.51	1.22	0.92	1.74	1.40	1.06
Pooled-label DR ($\mathcal{S}, \mathcal{L}$)	0.51	0.54	0.56	0.88	0.89	0.90	1.28	1.16	0.97	1.43	1.31	1.10
CAFE ($\mathcal{S}, \mathcal{L}$)	0.54	0.59	0.65	0.88	0.90	0.92	1.13	1.01	0.84	1.29	1.15	0.96

Table 10: Same-evaluator setting in Case 3. Columns give target-evaluation counts; parentheses in methods identify samples supplying observed scores, with $\mathcal{L} = \{j, R_j = 1\}$ indexing evaluated target tasks. Green/blue denote ranks 1/2.

Method	Full rank recovery ↑			Kendall's τ ↑			MAE ($\times 10^2$) ↓			RMSE ($\times 10^2$) ↓		
	200	300	500	200	300	500	200	300	500	200	300	500
Target-only sample mean ($\mathcal{L}$)	0.40	0.52	0.56	0.84	0.88	0.90	1.65	1.33	0.98	1.88	1.52	1.12
Source reweighting ($\mathcal{S}$)	0.67	0.67	0.67	0.92	0.92	0.92	1.88	1.88	1.88	2.04	2.04	2.04
PPI++ ($\mathcal{S}, \mathcal{L}$)	0.42	0.50	0.55	0.84	0.87	0.89	1.57	1.27	0.93	1.81	1.46	1.07
Target-only AIPW ($\mathcal{L}$)	0.39	0.49	0.58	0.83	0.87	0.90	1.63	1.31	0.94	1.87	1.50	1.07
Pooled-label AIPW ($\mathcal{S}, \mathcal{L}$)	0.39	0.50	0.57	0.83	0.87	0.89	1.55	1.26	0.91	1.80	1.44	1.05
Pooled-label DR ($\mathcal{S}, \mathcal{L}$)	0.63	0.67	0.69	0.92	0.93	0.94	1.54	1.41	1.21	1.68	1.54	1.32
CAFE ($\mathcal{S}, \mathcal{L}$)	0.64	0.71	0.75	0.92	0.94	0.95	1.26	1.14	0.92	1.39	1.26	1.03

E.3 The different-evaluator setting

E.3.1 Implementation Details of CAFE

As in the same-evaluator setting, $q = p_{\mathcal{T}}/p_{\mathcal{S}}$ is estimated by KMM using all source and target covariates, with the resulting source weights normalized to mean one and held fixed across folds. We also implement CAFE using five-fold cross-fitting. The regression m for benchmark-evaluator scores is fitted on the source and evaluated-target observations, while the regression $\widetilde{m}$ for target-evaluator scores is fitted on the evaluated-target observations only; both use ridge regression. Within each outer training fold, we choose $\lambda_a \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ by three-fold cross-validation, with target observations stratified by evaluation status. For each candidate value, the two regressions and the coefficient function are fitted on the inner training folds and evaluated using the held-out criterion corresponding to (6). We select the value with the lowest average validation criterion and refit the coefficient function using the resulting out-of-fold residuals from the full outer training sample.

We again model $a(x)$ in an RKHS with Gaussian kernel

$$k(x, x') = \exp\left\{-\frac{\|x - x'\|_2^2}{2\sigma^2}\right\},$$

where σ is set to the median pairwise Euclidean distance among at most 2,000 pooled source and target observations in the corresponding training sample. The target-evaluation fraction is computed from the data used to fit each coefficient function, while the final held-out scores use the overall empirical fraction $\widehat{\alpha} = n_{\mathcal{L}}/n_{\mathcal{T}}$.

E.3.2 Implementation Details of Baselines

For comparison, we include target-only sample mean, source reweighting, PPI++, RePPI, and target-only AIPW as baselines. Target-only sample mean, PPI++, and target-only AIPW follow

Appendix E.2.2, with target outcomes Y replaced by $\widetilde{Y}$ in both fitting and estimation; the source predictor $\widehat{f}_{\mathcal{S}}$ remains fitted on (X, Y) from $\mathcal{S}$. Unless otherwise specified, we estimate the density ratio $q = p_{\mathcal{T}}/p_{\mathcal{S}}$ by KMM using all source and target covariates and keep $\widehat{q}$ fixed across folds.

Source reweighting. Source reweighting uses the observed source scores together with source and target covariates. Since only benchmark-evaluator scores are available in the source sample, we transport these scores to the target distribution:

$$\widehat{\mu}_{\mathrm{SR}} = \frac{\mathbb{P}_{\mathcal{S}}[\widehat{q}(X)Y]}{\mathbb{P}_{\mathcal{S}}\widehat{q}(X)}.$$

RePPI. We use the three-fold mean-estimation procedure of Ji et al. (2025) and its public implementation[3]. Recalibration uses $(X, \widehat{f}_{\mathcal{S}}(X))$ as input. Let $\widehat{h}_\ell$ denote the recalibrated predictor in rotation ℓ and $\widehat{c}_\ell$ the fitted control-variate coefficient, including its evaluated/unevaluated sample-size adjustment. For the evaluated target fold $\mathcal{L}_\ell$,

$$\widehat{\mu}_{\mathrm{RePPI}} = \frac{1}{3}\sum_{\ell=1}^{3}\left[\mathbb{P}_{\mathcal{L}_\ell}\widetilde{Y} + \widehat{c}_\ell\left\{\mathbb{P}_{\mathcal{U}}\widehat{h}_\ell(X) - \mathbb{P}_{\mathcal{L}_\ell}\widehat{h}_\ell(X)\right\}\right].$$

The mean-specialized initialization cancels from this expression, although the initialization, recalibration, and evaluation folds still rotate. The recalibrated predictor uses only covariates available throughout the target sample.

E.3.3 Additional Results

Table 11: Different-evaluator setting in Case 1. Columns give target-evaluation counts; parentheses in methods identify samples supplying observed scores, with $\mathcal{L} = \{j, R_j = 1\}$ indexing evaluated target tasks. Green/blue denote ranks 1/2.

Method	Full rank recovery ↑			Kendall's τ ↑			MAE ($\times 10^2$) ↓			RMSE ($\times 10^2$) ↓		
	200	300	500	200	300	500	200	300	500	200	300	500
Target-only sample mean ($\mathcal{L}$)	0.50	0.56	0.60	0.89	0.91	0.92	1.67	1.36	1.04	1.91	1.56	1.18
Source reweighting ($\mathcal{S}$)	0.42	0.42	0.42	0.84	0.84	0.84	4.17	4.17	4.17	4.43	4.43	4.43
PPI++ ($\mathcal{S}, \mathcal{L}$)	0.48	0.57	0.61	0.88	0.91	0.92	1.61	1.31	1.00	1.85	1.50	1.14
RePPI ($\mathcal{S}, \mathcal{L}$)	0.48	0.56	0.58	0.88	0.91	0.92	1.65	1.34	1.00	1.90	1.54	1.14
Target-only AIPW ($\mathcal{L}$)	0.48	0.58	0.58	0.88	0.91	0.91	1.66	1.34	1.00	1.91	1.54	1.14
CAFE ($\mathcal{S}, \mathcal{L}$)	0.48	0.57	0.56	0.87	0.91	0.91	1.50	1.24	0.95	1.73	1.43	1.09

Table 12: Different-evaluator setting in Case 3. Columns give target-evaluation counts; parentheses in methods identify samples supplying observed scores, with $\mathcal{L} = \{j, R_j = 1\}$ indexing evaluated target tasks. Green/blue denote ranks 1/2.

Method	Full rank recovery ↑			Kendall's τ ↑			MAE ($\times 10^2$) ↓			RMSE ($\times 10^2$) ↓		
	200	300	500	200	300	500	200	300	500	200	300	500
Target-only sample mean ($\mathcal{L}$)	0.51	0.55	0.59	0.89	0.91	0.92	1.73	1.39	1.00	1.98	1.58	1.14
Source reweighting ($\mathcal{S}$)	0.66	0.66	0.66	0.92	0.92	0.92	4.34	4.34	4.34	4.53	4.53	4.53
PPI++ ($\mathcal{S}, \mathcal{L}$)	0.50	0.55	0.60	0.89	0.91	0.92	1.65	1.34	0.94	1.90	1.53	1.09
RePPI ($\mathcal{S}, \mathcal{L}$)	0.46	0.52	0.60	0.88	0.90	0.92	1.72	1.36	0.96	1.97	1.56	1.11
Target-only AIPW ($\mathcal{L}$)	0.49	0.54	0.60	0.88	0.90	0.92	1.72	1.36	0.96	1.97	1.55	1.11
CAFE ($\mathcal{S}, \mathcal{L}$)	0.52	0.55	0.62	0.90	0.91	0.92	1.54	1.31	0.98	1.77	1.49	1.12

[3] `https://github.com/Wenlong2000/RePPI`.