LABEL-EFFICIENT RESIDUAL CORRECTION FOR MIXTURE-OF-EXPERTS VIA THEIR ROUTED READOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained mixture-of-experts (MoE) predictors are increasingly reused on heterogeneous deployment data. In such settings, the model's error, whether caused by distribution shift or by limited training, can differ across the input regions served by different experts, yet retraining is often infeasible due to restricted access to data or model parameters. Existing post-hoc adaptation methods typically treat the model as a black-box predictor, ignoring the routing weights and expert outputs that form each prediction. We propose ATF-KNN (Affine Teacher Features with kNN residual imputation), a post-hoc adaptation method for a frozen MoE teacher, which leverages the routing probabilities and expert-specific outputs exposed at inference time. Our approach fits one offset and one gain per expert in one step, gated by the model's own routing probabilities, and imputes residuals for unlabeled deployment data via k-nearest neighbors in this readout space. We characterize the population gain over scalar teacher correction and prove a finite-sample risk certificate under regularity conditions. Experiments on real and synthetic data show consistent improvements over the frozen teacher under both limited training and distribution shift, robustness to a misspecified number of experts and over a range of imputation hyperparameters, and, given enough labels, gains over scalar correction when the teacher's residual depends on its routing.

1 INTRODUCTION

Pretrained predictors are increasingly reused on data whose distribution differs from that of their training data. A broad foundation-style model may need to serve a specialized task, a hospital-specific population, a new product domain, or a deployment population whose covariate distribution or conditional outcome distribution differs from that of the training data. In such settings the model's predictions remain predictive of the outcome but carry a systematic error. A small number of deployment labels can often be collected and many unlabeled deployment covariates may be available, yet revisiting the original data, weights, or training pipeline is unavailable or too expensive. This scarce-label adaptation regime has motivated a rapidly growing line of work on learning and inference from teacher predictions—prediction-powered inference (Angelopoulos et al., 2023a;b), post-prediction inference (Miao & Lu, 2024; Miao et al., 2025), online-tuned prediction-powered learning (Shoham et al., 2025), multi-predictor aggregation (Cowen-Breen et al., 2026; Shan et al., 2026), Bayesian priors on prediction quality (Cortinovis & Caron, 2025), and imputation-powered inference (Zhao & Candès, 2025). In these works the teacher's predictions are correlated with the outcome, and a small labeled sample from the target distribution corrects them or calibrates inference based on them. However, treating the teacher only as a source of scalar predictions discards how each prediction was formed, which matters when the model routes different inputs to different components.

Mixture-of-experts (MoE) predictors are designed for heterogeneous data (Jacobs et al., 1991; Yuksel et al., 2012): a gate routes each input to experts that specialize in different regimes, such as subpopulations, operating conditions, or source domains. This specialization shapes how a pretrained MoE fails: an expert that saw few examples of its regime, or that serves inputs whose distribution changes at deployment, can be biased while the other experts remain accurate. The MoE also exposes information that can locate this error: beyond the scalar prediction $\hat{f}(x) = \sum_k \pi_k(x)\mu_k(x)$, its forward pass computes a routing distribution $\pi(x) \in \Delta^{K-1}$ over the experts and the per-expert

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.

readouts $\mu_k(x)$. This suggests correcting a pretrained MoE by making full use of its structure: can a few labels correct each expert's error through these outputs, without retraining the model? Answering this requires only *gray-box* access: $\pi(x)$ and $\mu(x)$ are read from forward passes, while the model's parameters and training data are neither used nor modified, as they would be in the case of white-box fine-tuning. Such a correction should vary with the routing, so that each expert's error is corrected separately, while having few enough parameters to be estimated from the few labels available.

We focus on point regression $\mathbb{E}[Y \mid X]$ from a small labeled set $\mathcal{S}$, an auxiliary unlabeled covariate pool $\mathcal{U}$, and a given MoE teacher whose gate and per-expert outputs are readable but whose weights are neither used nor modified. A direct pseudo-labeling approach would use $\hat{f}(x_u)$ as labels for auxiliary covariates, while scalar post-prediction methods would treat $\hat{f}(x)$ as the only teacher-derived quantity. Both choices reduce the MoE to its scalar output and discard the routing weights and expert predictions, although the teacher's error may depend on them. ATF-KNN addresses this gap by learning a post-hoc residual update in the teacher's own readout coordinates while leaving the teacher unchanged. It lies between scalar correction, which uses only $\hat{f}(x)$, and fine-tuning, which updates the model parameters: it fits $2K$ coefficients, one offset and one gain per expert. Figure 1 summarizes this post-hoc adaptation workflow.

Our contributions are:

1. We introduce ATF-KNN, a one-step residual-adaptation estimator that combines affine MoE readout features, kNN residual imputation on unlabeled covariates, and a weighted ridge update for data-efficient expert-specific correction. This form corrects errors that differ across experts, yet it has only $2K$ coefficients, independent of the teacher's size, and is fitted in closed form without parameter access or iterative training, at a cost linear in the number of unlabeled covariates (Algorithm 1).
2. We provide a theoretical analysis of affine MoE readout correction, showing when projection onto teacher-feature coordinates can improve over scalar correction and deriving finite-sample conditions that quantify the label-efficiency trade-off.
3. We evaluate the method across synthetic and real benchmarks under IID, expert-number misspecification, and deployment-shift settings, showing broad gains from structured MoE readout over the given teacher, scalar teacher-use baselines, and matches or improves on fine-tuning with far fewer parameters.

2 Related Work

Scarce-label correction with pretrained predictors. Semi-supervised estimation with limited labels and many covariates has been studied through adjusted estimators and efficient regression procedures (Chakrabortty & Cai, 2018; Zhang et al., 2019; Azriel et al., 2022). Related prediction- and imputation-assisted inference methods use pretrained predictions as auxiliary information, including prediction-powered inference (Angelopoulos et al., 2023a;b), post-prediction inference (Miao & Lu, 2024; Miao et al., 2025), online-tuned prediction-powered learning (Shoham et al., 2025), multi-predictor aggregation (Cowen-Breen et al., 2026; Shan et al., 2026), Bayesian priors on prediction quality (Cortinovis & Caron, 2025), and imputation-powered inference (Zhao & Candès, 2025). These methods typically consume scalar predictions or low-dimensional black-box outputs. ATF-KNN instead uses the deterministic MoE readout $\phi(x) = [\pi_{1:K-1}, \pi \odot \mu]$ as the correction coordinates for point prediction.

Pseudo-labeling and self-training. Self-training, consistency regularization, confidence-filtered pseudo-labeling, and tri-training form the standard pseudo-labeling toolkit (Zhou & Li, 2005; Tarvainen & Valpola, 2017; Berthelot et al., 2019; Xie et al., 2020; Sohn et al., 2020); see van Engelen & Hoos (2020) for a survey and Oliver et al. (2018) for evaluation protocols in deep SSL. In regression, the direct analogue refits on $(\mathcal{S} \cup \mathcal{U},\ y_s \cup \hat{f}(x_u))$ and inherits systematic teacher bias. Our method imputes residuals rather than labels and forms neighborhoods in MoE-readout space.

Mixture-of-experts and routed state. MoE models originate in local-expert regression and have become central sparse routed neural architectures (Jacobs et al., 1991; Yuksel et al., 2012; Shazeer et al., 2017; Fedus et al., 2022). Recent work also uses MoE gates after training, for example as

a similarity signal for adaptive conformal prediction (Kong et al., 2026). Our setting is post-hoc point-regression adaptation: the teacher is fixed, the gate and expert readouts are available, and the correction is learned in their affine span. The kNN and ridge components rely on standard nearest-neighbor regression, ridge regularization, and matrix concentration tools (Györfi et al., 2002; Hoerl & Kennard, 1970; Tropp, 2012). The deployment-shift stress tests use the standard covariate-shift framing, where the covariate marginal changes while the conditional law is preserved (Sugiyama et al., 2007; Pan & Yang, 2010). The task is not transfer learning (Pan & Yang, 2010; Kuzborskij & Orabona, 2013): instead of training a predictor for a new target domain, ATF-kNN corrects the frozen teacher's residual on the deployment distribution, whether or not it stems from a shift.

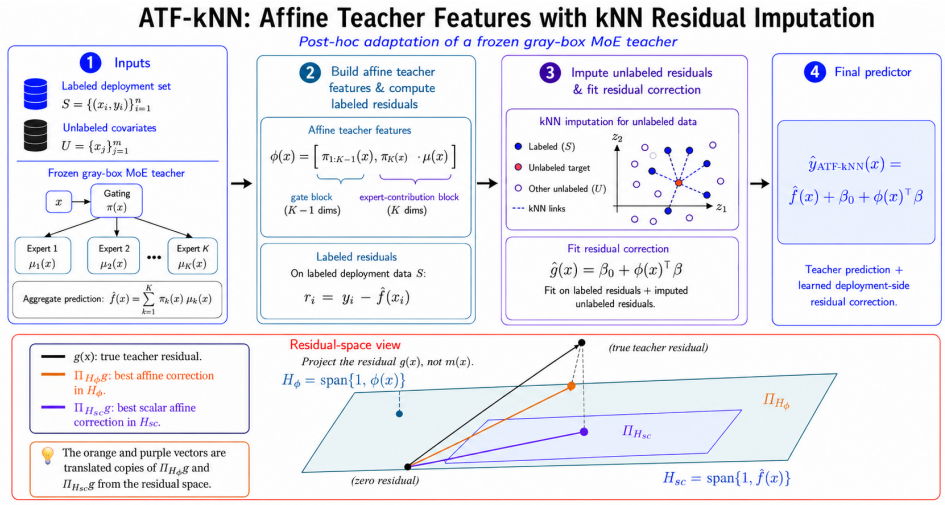


Figure 1: ATF-kNN uses frozen MoE routing and expert readouts to form affine teacher features, imputes residuals by kNN in this feature space, and fits a weighted ridge correction added to the teacher prediction.

3 Methodology

3.1 Deployment Setting and Predictor Readout

Let $(X, Y) \sim P$, where $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and $Y \in \mathbb{R}$. At deployment time, we observe labeled data $\mathcal{S} = \{(x_s^i, y_s^i)\}_{i=1}^{n_s}$ and unlabeled covariates $\mathcal{U} = \{x_u^j\}_{j=1}^{n_u}$ from the same deployment distribution; write $\mathcal{S}_X = \{x_s^i\}_{i=1}^{n_s}$ for the labeled covariates. The target is the conditional mean $m(x) = \mathbb{E}[Y \mid X = x]$, with squared prediction risk

$$\mathcal{R}(\tilde{m}) = \mathbb{E}[(\tilde{m}(X) - m(X))^2]. \tag{1}$$

We are also given a fixed gray-box MoE predictor, pretrained on data from a source distribution P_0 with regression function $m_0(x) = \mathbb{E}_{P_0}[Y \mid X = x]$. A forward pass at any x returns routing probabilities, expert readouts, and the aggregated prediction

$$\hat{f}(x) = \sum_{k=1}^{K} \pi_k(x)\mu_k(x), \qquad \pi(x) = (\pi_1(x), \ldots, \pi_K(x)) \in \Delta^{K-1}, \qquad \mu_k(x) \in \mathbb{R}. \tag{2}$$

Here K denotes the number of exposed experts, $\Delta^{K-1} = \{p \in \mathbb{R}^K : p_k \geq 0, \sum_{k=1}^K p_k = 1\}$ denotes the $(K-1)$-dimensional probability simplex, $\pi_k(x)$ is the routing weight for expert k, and $\mu_k(x)$ is that expert's scalar prediction. The parameters, pretraining data, and training pipeline of the MoE predictor are neither used nor modified. Therefore, we condition throughout on the realized functions $(\hat{f}, \pi, \mu)$, without assuming that the exposed gate or experts are correctly specified for the deployment distribution.

Remark 1 (Imperfect-teacher setting). The method is designed for teachers that are useful but systematically imperfect on P. With residual $g = m - \hat{f}$, the teacher is systematically imperfect if $\mathbb{E}[g(X)^2] > 0$, so that its error $Y - \hat{f}(X) = g(X) + \{Y - m(X)\}$ has a nonzero conditional mean. This occurs in two ways. First, $P = P_0$, but $\hat{f} \neq m_0$ because training was limited, for example by the amount of training data or the computing budget. Second, $P \neq P_0$: a change in the regression function adds $m - m_0$ to the residual, and a change in the covariate distribution moves probability mass to inputs where $\hat{f}$ is less accurate. In both cases, ATF-KNN uses a small labeled sample from P, as in prediction-powered and multi-predictor inference (Angelopoulos et al., 2023a;b; Cowen-Breen et al., 2026), to estimate the projection of g onto the span H_ϕ of a constant and the teacher's readout features (Section 3.2); the teacher is useful for correction when this projection is nonzero, and the population gain over the teacher is its squared norm (Theorem 1).

Remark 2 (Choice of the expert count K). In ATF-KNN, K is the number of experts exposed by the frozen teacher. Larger K enlarges the span of the readout features, so the correction can differ across more experts, but also increases the affine feature dimension $p_\phi = 2K - 1$ and may worsen finite-sample conditioning. Smaller K gives fewer coefficients and a smaller correction span. We study this practical trade-off empirically through an expert-number robustness analysis in Section 5.2.

The proposed ATF-KNN estimates the residual g on top of the teacher, $\widehat{m}(x) = \hat{f}(x) + \widehat{g}(x)$, where $\widehat{g}$ is a fitted residual correction. On labeled deployment points, the observable residuals are $r_s^i = y_s^i - \hat{f}(x_s^i)$ for $i = 1, \ldots, n_s$.

3.2 Affine Teacher-Feature kNN Method

A scalar correction depends on x only through $\hat{f}(x)$, so it can only recalibrate the final prediction, for example as $a + b\hat{f}(x)$, in the same way for all experts. In contrast, we consider the affine teacher-feature map. For each input, define the affine teacher-feature vector

$$\phi(x) = \big[\underbrace{\pi_1(x), \ldots, \pi_{K-1}(x)}_{\text{gate block}}, \underbrace{\pi_1(x)\mu_1(x), \ldots, \pi_K(x)\mu_K(x)}_{\text{expert-contribution block}}\big] \in \mathbb{R}^{p_\phi}, \qquad p_\phi = 2K - 1. \tag{3}$$

Using both routing probabilities and expert contributions allows the residual correction to depend on which experts are active and on their individual predictions. The last gate coordinate is excluded because $\sum_{k=1}^{K} \pi_k(x) = 1$; with an intercept in the residual model, retaining all K gate coordinates would introduce a linear dependence.

For $a \in \{s, u\}$, let $\Phi_a = [\phi(x_a^1), \ldots, \phi(x_a^{n_a})]^\top \in \mathbb{R}^{n_a \times p_\phi}$, where $n_s = |\mathcal{S}|$ and $n_u = |\mathcal{U}|$. Because gate coordinates and expert-contribution coordinates may have different numerical scales, all distances and regressions use the standardized features $\tilde{\phi}_\ell(x) = \phi_\ell(x)/\hat{s}_\ell$, where $\hat{s}_\ell$ is the standard deviation of coordinate ℓ on $\mathcal{S}_X \cup \mathcal{U}$; standardization needs no labels. The resulting matrices are denoted by $\widetilde{\Phi}_s$ and $\widetilde{\Phi}_u$, and $\tilde{\phi}(x_\star)$ denotes the standardized feature for a query point $x_\star$.

Step 1: Impute unlabeled residuals. For each unlabeled point x_u^j, compute its distance to labeled point x_s^i in standardized teacher-feature space:

$$d_{ij} = \left\| \tilde{\phi}(x_u^j) - \tilde{\phi}(x_s^i) \right\|_2, \quad i = 1, \ldots, n_s.$$

Let $\mathcal{N}_{k_{\text{nn}}}(j)$ be the indices of the k_{nn} nearest labeled points, where k_{nn} is a neighbor count and is distinct from the expert count K. ATF-KNN imputes the residual for x_u^j by

$$\hat{r}_u^j = \sum_{i \in \mathcal{N}_{k_{\text{nn}}}(j)} a_i^{(j)} r_s^i, \qquad a_i^{(j)} = \frac{(d_{ij} + \epsilon)^{-1}}{\sum_{\ell \in \mathcal{N}_{k_{\text{nn}}}(j)} (d_{\ell j} + \epsilon)^{-1}}, \tag{4}$$

with $\epsilon > 0$ used only to make the inverse-distance weights well-defined at zero distance. Residual information is then transferred between deployment points that induce similar gate probabilities and expert contributions, rather than between points that are close only in raw covariate space. Replacing $a_i^{(j)}$ in (4) by $1/k_{\text{nn}}$ gives the ordinary uniform kNN imputer on the same neighbor set.

Remark 3 (Role of unlabeled covariates). The unlabeled sample is used without responses. It contributes to the deployment-scale standardization of ϕ and supplies additional design points whose residuals are imputed from labeled neighbors in the same teacher-feature space. Algorithm 1 in the appendix summarizes these two uses.

Step 2: Fit the residual correction. After imputation, stack the residual targets and design matrix as

$$\widehat{r} = (r_s^1, \ldots, r_s^{n_s}, \hat{r}_u^{\,1}, \ldots, \hat{r}_u^{\,n_u})^\top, \qquad \widetilde{\Phi} = [\widetilde{\Phi}_s; \widetilde{\Phi}_u].$$

Labeled residuals receive weight 1, and imputed residuals receive weight w_u: $\omega_i = 1$ for $i \le n_s$ and $\omega_{n_s+j} = w_u \in [0,1]$ for $j \le n_u$. Since n_u can be much larger than n_s, the aggregate influence of the imputed block is governed by the effective unlabeled-to-labeled weight ratio $\rho_u = w_u n_u / n_s$. Letting $\widehat{r}_t$ denote the t-th element of the stacked residual vector $\widehat{r}$, the final residual model is the weighted ridge estimator, with an unpenalized intercept:

$$(\hat{\beta}, \hat{\beta}_0) = \underset{\beta \in \mathbb{R}^{p_\phi},\, \beta_0 \in \mathbb{R}}{\arg\min} \sum_{t=1}^{n_s+n_u} \omega_t \left(\widehat{r}_t - \beta_0 - \widetilde{\Phi}_{t\cdot}^\top \beta \right)^2 + \alpha \left\| \beta \right\|_2^2 . \tag{5}$$

In (5), $\widetilde{\Phi}_{t\cdot}$ denotes the t-th row of the stacked design matrix $\widetilde{\Phi}$; rows $1, \ldots, n_s$ correspond to labeled covariates and rows $n_s + j$ correspond to unlabeled covariates. For a query point $x_\star$, the resulting prediction is

$$\hat{y}_{\text{ATF-kNN}}(x_\star) = \hat{f}(x_\star) + \hat{\beta}_0 + \tilde{\phi}(x_\star)^\top \hat{\beta}. \tag{6}$$

From a geometric viewpoint, a linear residual model in ϕ is equivalent to a gate-weighted affine adjustment of the expert readouts:

$$\beta_0 + \phi(x)^\top \beta = \sum_{k=1}^{K} \pi_k(x) \{ a_k + b_k \mu_k(x) \},$$

with $a_K = \beta_0$, $a_k = \beta_0 + \beta_k$ for $k < K$, and $b_k = \beta_{K-1+k}$. Thus the gate block represents expert-specific offsets, and the expert-contribution block represents expert-specific gain adjustments. Since $\hat{f} = \sum_k \pi_k \mu_k$, the corrected predictor $\sum_k \pi_k(x) \{ a_k + (1 + b_k) \mu_k(x) \}$ is the same mixture, with the same routing, in which each expert's readout is recalibrated by one offset and one gain. The scalar correction $\hat{f}(x) + a$ is recovered by the constrained case $b_k = 0$ and $a_1 = \cdots = a_K = a$. From the projection viewpoint, the affine teacher-feature space $H_\phi = \operatorname{span}\{1, \phi(x)\}$ contains the scalar correction space $H_{\text{sc}} = \operatorname{span}\{1, \hat{f}(x)\}$. Hence the affine features improve on scalar correction exactly when the teacher residual has a nonzero component in H_ϕ orthogonal to H_{sc}, and the population gain is the squared L^2 norm of that component (Theorem 1; Figure 1).

Remark 4. Equation (5) uses the unnormalized ridge parameter α. In the finite-sample analysis, the corresponding normalized penalty is $\lambda = \alpha / W$, where $W = n_s + w_u n_u$ is the total sample weight. In the experiments, α is selected by labeled-set cross-validation over a fixed grid whenever the benchmark protocol does not prescribe a fixed default; the same grid is applied to every ridge-based method in the corresponding protocol. The penalty controls the variance of the finite-sample fit; the residual function class is determined by the affine MoE features in (3).

4 Theoretical Analysis

We condition throughout on the realized gray-box teacher; no result assumes correct gates, correct expert count, or unbiased experts. Let $m(x) = \mathbb{E}[Y \mid X = x]$, $g(x) = m(x) - \hat{f}(x)$, and $\mathcal{R}(\tilde{m}) = \mathbb{E}[(\tilde{m}(X) - m(X))^2]$. For the theory, let $Z(X) = D^{-1}\phi(X)$ be a population-scaled affine teacher-feature vector, with D positive diagonal, and define $\psi(X) = (1, Z(X)^\top)^\top \in \mathbb{R}^{2K}$ and $\Sigma = \mathbb{E}[\psi(X)\psi(X)^\top]$. Let $\mathcal{H}_\phi = \{x \mapsto \psi(x)^\top \theta : \theta \in \mathbb{R}^{2K}\}$ and $\mathcal{H}_{\text{sc}} = \operatorname{span}\{1, \hat{f}\}$. Since $\hat{f}(x) = \sum_k \pi_k(x) \mu_k(x)$ is a linear combination of the product block of ϕ, $\mathcal{H}_{\text{sc}} \subseteq \mathcal{H}_\phi$. For any finite-dimensional $\mathcal{H} \subset L^2(P_X)$, let $\Pi_{\mathcal{H}}$ be the $L^2(P_X)$ projection and set $m_{\mathcal{H}} = \hat{f} + \Pi_{\mathcal{H}} g$. We write $m_\phi = m_{\mathcal{H}_\phi}$, $m_{\text{sc}} = m_{\mathcal{H}_{\text{sc}}}$, $\mathcal{A}_\phi = \mathcal{R}(m_\phi)$, $\Delta_\phi = \mathcal{R}(\hat{f}) - \mathcal{R}(m_\phi)$, and $G_{\phi|\text{sc}} = \mathcal{R}(m_{\text{sc}}) - \mathcal{R}(m_\phi)$.

Assumption 1 (Regular readout and residual)**.** The residual satisfies $g \in L^2(P_X)$, and there exist constants $L < \infty$ and $\kappa > 0$ such that $\|\psi(X)\|_2 \le L$ almost surely and $\lambda_{\min}(\Sigma) \ge \kappa$.

Theorem 1 (Population gain and scalar comparison)**.** *Under Assumption 1,*

$$\mathcal{A}_\phi = \left\| (I - \Pi_{\mathcal{H}_\phi}) g \right\|_{L^2(P_X)}^2, \qquad \Delta_\phi = \left\| \Pi_{\mathcal{H}_\phi} g \right\|_{L^2(P_X)}^2, \qquad G_{\phi|\text{sc}} = \left\| \Pi_{\mathcal{H}_\phi \cap \mathcal{H}_{\text{sc}}^\perp} g \right\|_{L^2(P_X)}^2 .$$

Thus the improvement over the original teacher is the residual energy captured by the affine teacher-feature readout, and the additional improvement over scalar affine correction is the captured residual energy orthogonal to $\operatorname{span}\{1, \hat{f}\}$.

Appendix A.2 bounds $\mathcal{A}_\phi$ when the deployment regression function is a mixture of recalibrated experts under a possibly different gate, and Appendix A.3 bounds it when the conditional mean of the residual given the readout features is smooth.

Let $\widehat{m}_{\mathrm{ATF}}(x) = \hat{f}(x) + \psi(x)^\top \widehat{\theta}$ denote the population-scaled analogue of Algorithm 1.

Theorem 2 (Finite-sample correction certificate). *Let $\widehat{\Sigma}_w$ be the weighted empirical covariance of ψ over $\mathcal{S}_X \cup \mathcal{U}$. Suppose the covariance stability event $\lambda_{\min}(\widehat{\Sigma}_w) \geq \kappa/2$ holds. Under Assumption 1 and the residual-noise and kNN-imputation conditions stated in Appendix A.5, the following bounds hold with high probability:*

$$\mathcal{R}(\widehat{m}_{\mathrm{ATF}}) \leq \mathcal{A}_\phi + \varepsilon_{\mathrm{stat}}, \qquad \mathcal{R}(\hat{f}) - \mathcal{R}(\widehat{m}_{\mathrm{ATF}}) \geq \Delta_\phi - \varepsilon_{\mathrm{stat}},$$
$$\mathcal{R}(m_{\mathrm{sc}}) - \mathcal{R}(\widehat{m}_{\mathrm{ATF}}) \geq G_{\phi|\mathrm{sc}} - \varepsilon_{\mathrm{stat}}.$$

Here $\varepsilon_{\mathrm{stat}}$ is a nonnegative finite-sample error term consisting of ridge estimation error, labeled residual-noise error, kNN imputation error on the unlabeled block, and ridge shrinkage bias.

The explicit expression for $\varepsilon_{\mathrm{stat}}$ and the full probability statement are given in Appendix A.5. The theorem says that the fitted correction improves over the teacher or scalar correction when the corresponding population gain, Δ_ϕ or $G_{\phi|\mathrm{sc}}$, exceeds $\varepsilon_{\mathrm{stat}}$.

Proposition 3 (Empirical covariance stability). *Let $W = n_s + w_u n_u$ and $N_w = (n_s + w_u n_u)^2/(n_s + w_u^2 n_u)$, and let $\widehat{\Sigma}_w$ be the weighted empirical covariance of ψ over $\mathcal{S}_X \cup \mathcal{U}$ with labeled weight 1 and unlabeled weight w_u. Under Assumption 1 and $w_u \in [0, 1]$, there is a universal constant $c > 0$ such that, if $W \geq c(L^2/\kappa)\log(16K/\delta)$ and*

$$N_w \geq c\,\frac{L^2\,\|\Sigma\|_{\mathrm{op}}}{\kappa^2}\log\frac{16K}{\delta},$$

then $\lambda_{\min}(\widehat{\Sigma}_w) \geq \kappa/2$ with probability at least $1 - \delta/4$. In the labeled-only case $w_u = 0$, $W = N_w = n_s$.

Proposition 3 gives a high-probability sufficient condition for the covariance stability event used in Theorem 2. The required effective sample size is linear in the $2K$ fitted readout coefficients up to conditioning and logarithmic factors (Remark 7). The expert count K is the number exposed by the realized teacher, not a true latent number of regimes; misspecifying it changes $\mathcal{A}_\phi$ and the constants L and κ, while the guarantees continue to hold conditionally on the realized teacher. The kNN step adds a local bias–variance trade-off in teacher-feature space; Appendix A.4 gives the weighted-kNN statement used in the proof.

Corollary 4 (Identifiability scale for affine readout correction). *In the labeled-only limit $w_u \to 0$, so that $W = N_w = n_s$, the covariance-stability requirement in Proposition 3 reduces, up to universal constants, to $n_s \gtrsim (L^2\,\|\Sigma\|_{\mathrm{op}}/\kappa^2)\log(16K/\delta)$. Thus the affine readout correction requires enough labeled information to identify a $2K$-dimensional residual model. This suggests the heuristic switch: lower-dimensional scalar correction can be competitive when $n_s \lesssim c^\star(2K-1)$, where $c^\star$ absorbs conditioning and logarithmic factors, while affine teacher-feature correction is expected to become preferable once n_s exceeds this scale and $G_{\phi|\mathrm{sc}}$ exceeds $\varepsilon_{\mathrm{stat}}$ (Theorem 2).*

5 Experiments

We evaluate ATF-kNN on real data under a natural temporal shift and with scarce labels (Section 5.1), and on controlled benchmarks: a cross-setting comparison and IID sweeps, expert-number misspecification, hyperparameter sensitivity, and deployment shift (Section 5.2). Appendix C reports a geographic shift and a synthetic suite whose shifts are fixed before the teacher is trained. The baseline set tests three alternatives to the proposed residual-imputation objective: scalar-only use of the teacher, supervised or pseudo-labeled fits that ignore the MoE readout, and MoE-readout adapted comparators based on prediction-powered SSL (Shoham et al., 2025) and SADA-style estimating equations (Shan et al., 2026), which separate the value of exposing (π, μ) from the value of fitting residuals with kNN-imputed unlabeled targets (Appendix F). As white-box controls, which update the teacher's parameters, we report L2-SP fine-tuning (Li et al., 2018) of the router, the expert heads, or the full model (FT in the tables). All methods share the labeled, unlabeled, and evaluation samples within each repetition, and paired comparisons use Holm correction (Holm, 1979); protocols are in Appendix B. Appendix E.2 adds a Fashion-MNIST probability-estimation experiment.

Table 1: **Electricity**, log-RMSE at $n_s = 40$ (30 repetitions); wins of ATF-KNN. †: ATF-KNN significantly better (Wilcoxon test, Holm).

Method	log-RMSE	ATF-KNN wins	Params
Teacher	0.4434†	29/30	0
Mean correction	0.4480†	29/30	1
Scalar affine	0.4118†	27/30	2
Target-only	0.5514†	30/30	17
PP-SSL	0.5926†	30/30	8
SADA	0.8856†	30/30	8
Router FT	0.4154†	28/30	1,732
Expert-head FT	0.4006	16/30	132
ATF-NO-KNN	0.3954	8/30	8
ATF-KNN	0.3993	–	8

Table 2: **Bike-Sharing**, RMSE (30 repetitions). †: as in Table 1, within each budget.

Method	n_s=10	n_s=20	n_s=40
Teacher	1.265†	1.265†	1.265†
Mean correction	1.231†	1.202†	1.185†
Target-only	1.523†	1.487†	1.459†
PP-SSL	1.365†	1.150	1.049
SADA	1.274	1.108	1.067
Router FT	1.215	1.172	1.107
Expert-head FT	1.265†	1.185†	1.107†
Full FT	1.225	1.166	1.074
ATF-NO-KNN	1.396†	1.258	1.078
ATF-KNN	1.179	1.119	1.070

5.1 Real-Data Benchmarks

Electricity: temporal shift. We predict `log1p` hourly consumption for 128 clients of the Electricity Load Diagrams data (Trindade, 2015). A four-expert MoE is trained on 2012, early-stopped on 2013, frozen, and adapted to 2014 from $n_s = 40$ labels and 6,000 unlabeled examples over 30 repetitions. ATF-KNN reduces log-RMSE from 0.4434 to 0.3993 (Table 1) and outperforms scalar affine correction (0.4118; better in 27 of 30 repetitions) and router fine-tuning (0.4154; 28 of 30), both significant after Holm correction. It is better than the readout-adapted PP-SSL and SADA in all 30 repetitions, and it is statistically tied with expert-head fine-tuning (0.4006) while fitting 8 coefficients instead of 132. ATF-NO-KNN is similarly accurate (0.3954): at this budget the gain comes from the readout correction, not kNN imputation.

Bike-Sharing: scarce labels with white-box controls. Bike-Sharing (Fanaee-T & Gama, 2014) with a frozen five-expert neural MoE, whose pretraining data retain only 25%, 12%, and 5% of three of the five hour-and-weather regimes, is our most label-constrained setting: at $n_s = 10$, the ten correction coefficients are estimated from ten labels. Over 30 repetitions, ATF-KNN attains the lowest RMSE at $n_s = 10$ (1.179; Table 2) and is significantly better than the teacher (1.265), mean correction, PP-SSL, and expert-head fine-tuning; here kNN imputation reduces RMSE by 0.2165 relative to ATF-NO-KNN. At $n_s = 20$ and 40 it is statistically tied with the best method and remains significantly more accurate than expert-head fine-tuning.

Routing diagnostics. Two diagnostics show that the routing carries correction signal. On Electricity, the teacher's mean residual differs across routing regions (permutation test, $p < 0.001$), and none of 200 random shuffles of the routing weights attains a lower error than the learned routing on held-out source and target data (Appendix C.4). In the synthetic suite of Appendix C.3, permuting the readout coordinates other than $\hat{f}$ jointly across examples, which leaves the scalar prediction unchanged, raises the RMSE of ATF-NO-KNN from 0.7503 to 0.8505 at $n_s = 40$ (worse in 26 of 30 worlds).

5.2 Controlled Benchmarks

Teacher-corruption patterns (gate mixing, expert bias, expert noise, pretrain feature masking) probe the same imperfect-prediction setting that motivates prediction-powered and semi-supervised correction methods (Angelopoulos et al., 2023a;b; Miao & Lu, 2024; Miao et al., 2025; Shoham et al., 2025; Shan et al., 2026); exact numerical settings, held-out evaluation sets, and CV protocols are in Appendices E.1 and E.3. Appendix D ablates the feature map and expands the baseline set.

IID benchmarks. Table 3 compares the methods on a synthetic corrupted-oracle benchmark with a linear MoE teacher (K=3) and on a synthetic teacher whose residual is only partially affine in ϕ; ATF-KNN has the lowest mean RMSE in both. On the corrupted oracle, it reduces RMSE from 0.599 to 0.397, ahead of MoE-adapted PP-SSL (0.462) and SADA (0.636). For the partially affine

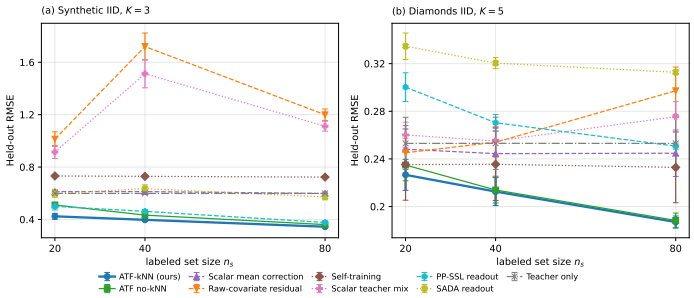


Figure 2: **IID benchmarks.** Held-out RMSE vs. n_s with ± 1 SE. *Left:* synthetic corrupted-oracle, linear MoE, K=3. *Right:* Diamonds, linear MoE, K=5, with pretraining subsampling and expert corruption. ATF-KNN is best for every n_s.

Table 3: **Cross-setting comparison**, $n_s = 40$. Held-out RMSE, mean $\pm$ std over 12 seeds.

Method	Synthetic (corrupted oracle)	Synthetic (non-affine)
Teacher	0.599 ± 0.004	0.330 ± 0.005
Mean correction	0.613 ± 0.045	0.327 ± 0.082
Self-training	0.729 ± 0.006	0.580 ± 0.007
PP-SSL	0.462 ± 0.040	0.352 ± 0.103
SADA	0.636 ± 0.101	0.475 ± 0.118
ATF-KNN	0.397 ± 0.044	0.310 ± 0.094

residual, the gain is smaller but positive (0.330 to 0.310), as expected when only the component of the nonlinear residual lying in $\mathrm{span}(\phi)$ is recoverable. In Figure 2, ATF-KNN has the lowest mean for every IID n_s, reducing teacher RMSE by 29%–42% on synthetic and 10%–26% on Diamonds with a linear MoE teacher (K=5). ATF-NO-KNN is the closest ablation: its RMSE is 4%–20% higher on synthetic and nearly matches ATF-KNN on Diamonds, and the gap shrinks with n_s, suggesting that kNN imputation mainly helps when labeled residuals are not sufficient to estimate the affine correction. Scalar, raw-covariate, self-training, and scalar teacher-mix baselines do not close this gap: per-expert bias and gain errors are representable as routed offset/gain corrections in the $(2K-1)$-dimensional $\mathrm{span}(\phi)$, but not as a global scalar correction.

Misspecified expert number. We fix $K_{\text{true}} = 3$ and sweep $K_{\text{teacher}} \in \{2, 3, 4, 5\}$ under the synthetic shift setup. ATF-KNN has the lowest mean RMSE for every K_{teacher}, with a 15.6%–23.7% margin over the best non-ATF-KNN method (Table 4). Given a frozen teacher, ATF-KNN uses the exposed gates and expert readouts as correction coordinates rather than estimates of the latent mixture components; larger K_{teacher} enlarges the correction span but raises the feature dimension to $2K_{\text{teacher}} - 1$.

Hyperparameter sensitivity. On the synthetic corrupted-oracle benchmark at $n_s = 40$ (Table 5), moderate neighbor counts are stable: $k_{\text{nn}} = 5$ and 10 are essentially tied, and $k_{\text{nn}} \in \{3, 15, 20\}$ still improves over ATF-NO-KNN; the one-neighbor imputer is too noisy. The error decreases smoothly as the effective unlabeled weight $\rho_u = w_u n_u / n_s$ moves away from zero and saturates around $\rho_u = 5$ to 10; the default $w_u = 0.5$ ($\rho_u = 62.5$) lies on this plateau.

Controlled deployment shift. Under synthetic covariate shift, where source and deployment $p(X)$ differ while $p(Y \mid X)$ is preserved (Sugiyama et al., 2007; Pan & Yang, 2010), ATF-KNN is best at $n_s = 40, 80$ (Figure 3, left). At $n_s = 20$, scalar mean correction is lower by 2.9%, and paired tests mark ATF-KNN as statistically tied with the best mean. This matches the finite-sample trade-off of Corollary 4: ATF-KNN fits a $2K$-coefficient readout correction, while scalar correction fits one or two coefficients. Under a Diamonds stress whose deployment outcomes add regime-specific biases to the source regression function (right), ATF-KNN has the lowest mean at both label sizes, while

Table 4: **Expert-number misspecification**, synthetic shift, $K_{\text{true}} = 3$, $n_s = 40$; mean RMSE over 20 seeds. Standard deviations: Table 15.

Method	K_{teacher} 2	3	4	5
Teacher	1.222	1.129	0.901	0.909
Self-training	1.206	1.118	0.912	0.924
Scalar teacher mix	1.218	1.252	1.180	1.161
PP-SSL	1.081	1.086	0.885	1.047
SADA	1.551	1.448	1.185	1.472
ATF-KNN	0.912	0.828	0.705	0.695

Table 5: **Hyperparameter sensitivity**, synthetic corrupted oracle, $n_s = 40$; mean RMSE over 12 seeds, varying k_{nn} or $\rho_u = w_u n_u / n_s$. *: default. Standard errors: Tables 17 and 18.

k_{nn}	RMSE	ρ_u	RMSE
1	0.4420	0	0.4329
3	0.4076	0.25	0.4185
5*	0.3972	0.5	0.4117
10	0.3976	1	0.4051
15	0.4034	2	0.4005
20	0.4116	5	0.3978
ATF-NO-KNN	0.4329	10	0.3972
Teacher	0.5986	62.5*	0.3972

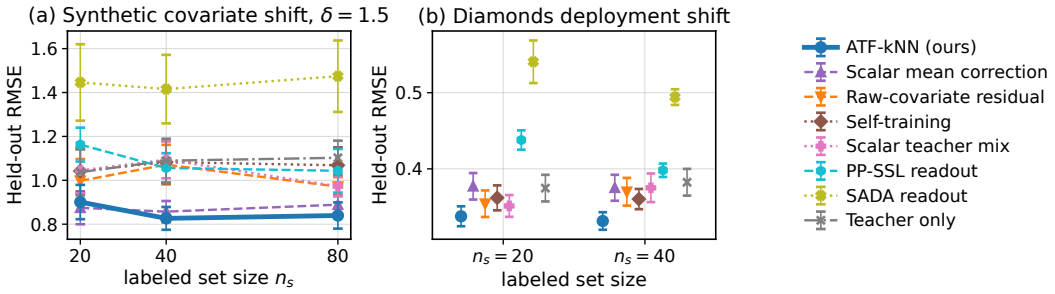


Figure 3: **Deployment-shift robustness.** *Left:* synthetic clean covariate shift, RMSE vs. n_s, ± 1 SE. *Right:* Diamonds with regime-specific outcome biases at deployment.

MoE-adapted PP-SSL and SADA have RMSE above the teacher, suggesting that readout coordinates alone are not enough without directly modeling residuals in that space.

Overall, ATF-KNN has the lowest mean on both IID sweeps, both settings of Table 3, the full expert-number misspecification sweep, and four of five deployment-shift cells; it is statistically tied in the $n_s = 20$ synthetic shift case. On both real-data benchmarks, it outperforms the simplest adaptation, scalar correction, and matches or improves on the white-box fine-tuning, while fitting only $2K$ coefficients. In short, by reading the routing weights and expert outputs from an MoE, ATF-KNN uses a few labels for an accurate correction efficiently, up to 90 times faster than fine-tuning (Table 10). Our method thus makes post-hoc correction accurate, label-efficient, and computationally light.

6 Discussion and Conclusion

ATF-KNN provides a lightweight gray-box adaptation for frozen MoE predictors. Rather than fine-tuning the teacher weights, it builds features $\phi(x) = [\pi_{1:K-1}(x), \pi(x) \odot \mu(x)]$, imputes residuals by kNN in standardized readout space, and fits a weighted ridge correction. Empirically, it is the best on the IID sweeps, the non-affine synthetic benchmark, and the expert-number misspecification sweep, and is the best or statistically tied under controlled deployment shift and on the Bike-Sharing neural-MoE benchmark. Theorems 1–2 formalize this behavior: the exposed readout can capture routed residual directions beyond scalar correction, and the fitted update has a finite-sample certificate under the stated regularity conditions without requiring correct latent K, correct gates, or unbiased experts.

The method requires gate probabilities and expert readouts, so it does not apply to scalar-only prediction interfaces. Scaling to very large sparse MoEs may require structured regularization, compression, learned metrics, or nonlinear residual models; multi-class/logit-space correction and joint fine-tuning are natural extensions. Overall, MoE forward passes expose adaptation signal beyond the scalar prediction, and ATF-KNN converts it into a label-anchored residual correction.

AI USE STATEMENT

We used generative AI tools to polish and aid the writing of this paper, including drafting and editing passages for clarity, checking the consistency of notation and terminology, and suggesting the title and keywords, and for retrieval and discovery, including finding related work and verifying and formatting references. Generative AI tools also assisted in verifying the proofs and, through an image-generation model, drew the workflow illustration in Figure 1 from our specification. We have reviewed all AI-assisted work: the authors checked the proof and all text, the table and figure scripts re-derive every reported statistic from the stored outputs, each reference was verified against its publication record, and Figure 1 was checked against the method description. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The ACSIncome data are public, de-identified census microdata that include demographic attributes; we use them only to evaluate a statistical correction method and make no decisions about individuals. The Electricity Load Diagrams data record anonymized consumption of electricity clients. The other data sets are public benchmarks or synthetic. We are not aware of other ethical concerns.

REPRODUCIBILITY STATEMENT

Complete proofs and all assumptions are in Appendix A. Data processing, teacher construction, label budgets, and hyperparameter grids for every experiment are in Appendices B and E, baseline implementations in Appendix F, and additional results in Appendix C. The supplementary material contains the code, configuration files, and per-repetition outputs for all reported experiments.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

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

David Azriel, Lawrence D. Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *Journal of the American Statistical Association*, 117(540): 2238–2251, 2022.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. MixMatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

Stefano Cortinovis and François Caron. FAB-PPI: Frequentist, assisted by Bayes, prediction-powered inference. In *International Conference on Machine Learning (ICML)*, 2025.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W. Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. Multiple-prediction-powered inference. In *International Conference on Learning Representations (ICLR)*, 2026.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 6478–6490, 2021.

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2:113–127, 2014.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Jingsen Kong, Wenlu Tang, Dezheng Kong, Linglong Kong, Guangren Yang, and Bei Jiang. Adaptive conformal prediction via mixture-of-experts gating similarity. In *International Conference on Learning Representations (ICLR)*, 2026.

Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *International Conference on Machine Learning (ICML)*, volume 28 of *Proceedings of Machine Learning Research*, pp. 942–950, 2013.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2825–2834, 2018.

Jiacheng Miao and Qiongshi Lu. Task-agnostic machine-learning-assisted inference. In *Advances in Neural Information Processing Systems 37*, 2024.

Jiacheng Miao, Xinran Miao, Yixuan Wu, Jiwei Zhao, and Qiongshi Lu. Assumption-lean and data-adaptive post-prediction inference. *Journal of Machine Learning Research*, 26(179):1–31, 2025.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Jiawei Shan, Zhifeng Chen, Yiming Dong, Yazhen Wang, and Jiwei Zhao. SADA: Safe and adaptive aggregation of multiple black-box predictions in semi-supervised learning. *The Annals of Applied Statistics*, 2026. arXiv:2509.21707.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Noa Shoham, Ron Dorfman, Shalev Shaer, Kfir Y. Levy, and Yaniv Romano. Prediction-powered semi-supervised learning with online power tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin D. Cubuk, Alex Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30*, 2017.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. URL https://doi.org/10.24432/C58C86.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Seniha Esen Yuksel, Joseph N. Wilson, and Paul D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012.

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.

Sarah Zhao and Emmanuel Candès. Imputation-powered inference. *arXiv preprint arXiv:2509.13778*, 2025.

Zhi-Hua Zhou and Ming Li. Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541, 2005.

APPENDIX

This appendix contains proofs, additional empirical evidence, and implementation details for ATF-KNN. Appendix A proves the population and finite-sample results from Section 4. Appendix B gives the protocols of the natural-shift, synthetic-suite, and Bike-Sharing experiments, and Appendix C their additional results, adaptation costs, and routing diagnostics. Appendix D reports feature-map ablations and additional baseline comparisons. Appendix E gives benchmark configurations, the Fashion-MNIST classification support experiment, and hyperparameter protocols. Appendix F specifies the baseline implementations, and Appendix G summarizes the reproduction protocol.

Algorithm 1 ATF-KNN — affine teacher-feature residual adaptation.

Require: Fixed MoE readout $(\hat{f}, \pi, \mu)$, labeled sample $\mathcal{S}$, unlabeled covariates $\mathcal{U}$, query point $x_\star$, hyperparameters $\alpha, k_{\mathrm{nn}}, w_u$.

1: Compute $\phi(x)$ from (3) for all $x \in \mathcal{S}_X \cup \mathcal{U} \cup \{x_\star\}$.
2: Estimate scales $\hat{s}_\ell$ on $\mathcal{S}_X \cup \mathcal{U}$ and form standardized features $\tilde{\phi}_\ell(x) = \phi_\ell(x)/\hat{s}_\ell$.
3: Form standardized feature matrices $\widetilde{\Phi}_s$ and $\widetilde{\Phi}_u$.
4: Compute labeled residuals $r_s^i \leftarrow y_s^i - \hat{f}(x_s^i)$ for $i = 1, \ldots, n_s$.
5: **for** $j = 1, \ldots, n_u$ **do**
6: Find $\mathcal{N}_{k_{\mathrm{nn}}}(j)$ among $\{\tilde{\phi}(x_s^i)\}_{i=1}^{n_s}$ using distance to $\tilde{\phi}(x_u^j)$.
7: Impute $\hat{r}_u^j$ by (4).
8: **end for**
9: Stack $\widetilde{\Phi} = [\widetilde{\Phi}_s; \widetilde{\Phi}_u]$, targets $\widehat{r} = (r_s^1, \ldots, r_s^{n_s}, \hat{r}_u^1, \ldots, \hat{r}_u^{n_u})^\top$, and weights $\omega = (\mathbf{1}_{n_s}, w_u \mathbf{1}_{n_u})^\top$.
10: Fit $(\hat{\beta}, \hat{\beta}_0)$ by (5).
11: **return** $\hat{y}_{\text{ATF-KNN}}(x_\star)$ from (6).

A ADDITIONAL THEORY AND PROOFS

This appendix proves the theory results from Section 4. All expectations are over the target distribution unless stated otherwise, and the teacher readout $(\hat{f}, \pi, \mu)$ is conditioned on throughout. For finite-sample statements, the labeled and unlabeled covariates are independent draws from the target covariate distribution, labeled responses are conditionally independent given their covariates, and the analysis uses the population-scaled feature $Z = D^{-1}\phi$ with $\psi = (1, Z^\top)^\top$. The unlabeled weight satisfies $w_u \in [0, 1]$ as in the method definition.

A.1 PROJECTION GEOMETRY

Lemma 5 (Risk identity for a projected residual). *Let $\mathcal{H} \subset L^2(P_X)$ be a finite-dimensional linear space, and let $\Pi_{\mathcal{H}}$ denote the $L^2(P_X)$ projection onto $\mathcal{H}$. For $m_{\mathcal{H}} = \hat{f} + \Pi_{\mathcal{H}} g$,*

$$\mathcal{R}(m_{\mathcal{H}}) = \|(I - \Pi_{\mathcal{H}})g\|^2_{L^2(P_X)}, \qquad \mathcal{R}(\hat{f}) - \mathcal{R}(m_{\mathcal{H}}) = \|\Pi_{\mathcal{H}} g\|^2_{L^2(P_X)}.$$

Proof. Since $g = m - \hat{f}$,

$$m_{\mathcal{H}}(X) - m(X) = \hat{f}(X) + \Pi_{\mathcal{H}} g(X) - m(X) = -(I - \Pi_{\mathcal{H}})g(X).$$

This gives the first identity. Also $\mathcal{R}(\hat{f}) = \mathbb{E}[g(X)^2] = \|g\|^2_{L^2(P_X)}$. The orthogonal decomposition $g = \Pi_{\mathcal{H}} g + (I - \Pi_{\mathcal{H}})g$ gives

$$\|g\|_2^2 - \|(I - \Pi_{\mathcal{H}})g\|_2^2 = \|\Pi_{\mathcal{H}} g\|_2^2,$$

which proves the second identity. □

Lemma 6 (Nested correction spaces). *With the affine teacher-feature map $\phi = [\pi_{1:K-1}, \pi \odot \mu]$, the scalar affine-teacher correction space satisfies*

$$\mathcal{H}_{\mathrm{sc}} \subseteq \mathcal{H}_\phi.$$

Furthermore, if $\mathcal{H}_0 \subseteq \mathcal{H}_\phi$, then

$$\mathcal{R}(m_{\mathcal{H}_0}) - \mathcal{R}(m_\phi) = \left\| \Pi_{\mathcal{H}_\phi \cap \mathcal{H}_0^\perp} g \right\|^2_{L^2(P_X)}.$$

Proof. For the inclusion, use

$$\hat{f}(x) = \sum_{k=1}^{K} \pi_k(x)\mu_k(x),$$

which is a linear combination of the product coordinates of ϕ. Together with the intercept, this places both 1 and $\hat{f}$ in $\mathcal{H}_\phi$.

For the risk difference, Lemma 5 gives

$$\mathcal{R}(m_{\mathcal{H}_0}) - \mathcal{R}(m_\phi) = \|(I - \Pi_{\mathcal{H}_0})g\|_2^2 - \|(I - \Pi_{\mathcal{H}_\phi})g\|_2^2.$$

Since $\mathcal{H}_0 \subseteq \mathcal{H}_\phi$,

$$\mathcal{H}_\phi = \mathcal{H}_0 \oplus (\mathcal{H}_\phi \cap \mathcal{H}_0^\perp)$$

after removing linearly dependent directions. Therefore,

$$\Pi_{\mathcal{H}_\phi} g = \Pi_{\mathcal{H}_0} g + \Pi_{\mathcal{H}_\phi \cap \mathcal{H}_0^\perp} g,$$

and the Pythagorean identity gives the displayed formula. □

Proof of the projection identities in Theorem 1. Assumption 1 gives $g \in L^2(P_X)$, bounded ψ, and $\lambda_{\min}(\Sigma) > 0$, so the quadratic objective

$$\mathbb{E}[(g(X) - \psi(X)^\top \theta)^2]$$

is strictly convex in θ. Differentiating gives the normal equation

$$\Sigma\theta^\star = \mathbb{E}[\psi(X)g(X)].$$

The identities for $\mathcal{R}(m_\phi)$ and Δ_ϕ follow from Lemma 5 with $\mathcal{H} = \mathcal{H}_\phi$. The comparison with any nested correction space follows from Lemma 6; taking $\mathcal{H}_0 = \mathcal{H}_{\mathrm{sc}}$ gives the scalar-comparison identity. □

Lemma 7 (Residualized-coordinate formula)**.** *Let $s(X)$ be any coordinate vector spanning a subspace $\mathcal{H}_0 \subseteq \mathcal{H}_\phi$, and let $a(X)$ be additional coordinates such that* $\mathrm{span}\{s, a\} = \mathcal{H}_\phi$*. Define*

$$\widetilde{a}(X) = a(X) - \mathbb{E}[a(X)s(X)^\top]\{\mathbb{E}[s(X)s(X)^\top]\}^\dagger s(X),$$

$$\Gamma = \mathbb{E}[\widetilde{a}(X)\widetilde{a}(X)^\top], \qquad b = \mathbb{E}[\widetilde{a}(X)g(X)].$$

Then

$$\mathcal{R}(m_{\mathcal{H}_0}) - \mathcal{R}(m_\phi) = b^\top \Gamma^\dagger b.$$

Proof. By construction, each coordinate of $\widetilde{a}$ is orthogonal in $L^2(P_X)$ to $\mathcal{H}_0$. Its span is the orthogonal complement of $\mathcal{H}_0$ inside $\mathcal{H}_\phi$, up to linearly dependent directions. The projection of g onto this residualized span has minimum-norm coefficient $\Gamma^\dagger b$, and the squared norm of the projection is $b^\top \Gamma^\dagger b$. Lemma 6 identifies this squared norm with $\mathcal{R}(m_{\mathcal{H}_0}) - \mathcal{R}(m_\phi)$. □

A.2 Imperfect Expert and Gate Decomposition

Proposition 8 (Routed approximation and gain lower bound)**.** *Suppose, for interpretation, that $m(x) = \sum_k \pi_k^\star(x)\nu_k(x)$ for a reference gate $\pi^\star$ and reference experts ν_k with $|\nu_k(X)| \le M_\nu$. For any constants $(a_k, b_k)_{k=1}^K$, write*

$$\eta_{k,a,b}(x) = \nu_k(x) - a_k - b_k\mu_k(x),$$

$$c_{a,b}(x) = \sum_{k=1}^{K} \pi_k(x)\{a_k + (b_k - 1)\mu_k(x)\},$$

and

$$s_{a,b}(x) = \sum_{k=1}^{K} \pi_k(x)\eta_{k,a,b}(x) + \sum_{k=1}^{K} \{\pi_k^\star(x) - \pi_k(x)\}\nu_k(x).$$

Then $c_{a,b} \in \mathcal{H}_\phi$ *and*

$$\mathcal{A}_\phi \le \inf_{a,b} \mathbb{E}[s_{a,b}(X)^2], \qquad \Delta_\phi \ge \sup_{a,b} \left(\|c_{a,b}\|_{L^2(P_X)} - \|s_{a,b}\|_{L^2(P_X)} \right)_+^2 .$$

In particular, if

$$E_{\text{aff}} := \inf_{a,b} \mathbb{E}\left[\sum_k \pi_k(X)\{\nu_k(X) - a_k - b_k\mu_k(X)\}^2 \right], \qquad E_{\text{gate}} := \mathbb{E}\,\|\pi^\star(X) - \pi(X)\|_1^2 ,$$

then $\mathcal{A}_\phi \le 2E_{\text{aff}} + 2M_\nu^2 E_{\text{gate}}$.

Proof. Start from

$$g(x) = m(x) - \hat{f}(x) = \sum_{k=1}^{K} \pi_k^\star(x)\nu_k(x) - \sum_{k=1}^{K} \pi_k(x)\mu_k(x).$$

Add and subtract $\sum_k \pi_k(x)\nu_k(x)$:

$$g(x) = \sum_{k=1}^{K} \pi_k(x)\{\nu_k(x) - \mu_k(x)\} + \sum_{k=1}^{K} \{\pi_k^\star(x) - \pi_k(x)\}\nu_k(x).$$

Using $\nu_k = a_k + b_k\mu_k + \eta_{k,a,b}$ gives

$$\begin{aligned} g(x) = &\sum_{k=1}^{K} \pi_k(x)\{a_k + (b_k - 1)\mu_k(x)\} \\ &+ \sum_{k=1}^{K} \pi_k(x)\eta_{k,a,b}(x) + \sum_{k=1}^{K} \{\pi_k^\star(x) - \pi_k(x)\}\nu_k(x), \end{aligned}$$

which is $c_{a,b}(x) + s_{a,b}(x)$.

The term $c_{a,b}$ belongs to $\mathcal{H}_\phi$. Indeed,

$$\sum_{k=1}^{K} a_k\pi_k(x) = a_K + \sum_{k=1}^{K-1} (a_k - a_K)\pi_k(x),$$

which is an intercept plus dropped-last gate coordinates, and

$$\sum_{k=1}^{K} (b_k - 1)\pi_k(x)\mu_k(x)$$

is a linear combination of the product coordinates. Therefore, the best ATF projection error is no larger than the error from choosing $c_{a,b}$ as the correction:

$$\mathcal{A}_\phi \le \mathbb{E}[(g(X) - c_{a,b}(X))^2] = \mathbb{E}[s_{a,b}(X)^2].$$

Taking the infimum over a, b gives the first upper bound.

For the lower bound on Δ_ϕ, if $\|c_{a,b}\|_2 = 0$ the claimed lower bound is zero and is immediate. Otherwise, since $c_{a,b} \in \mathcal{H}_\phi$,

$$\left\|\Pi_{\mathcal{H}_\phi} g\right\|_2 \ge \frac{\langle g, c_{a,b}\rangle}{\|c_{a,b}\|_2} = \|c_{a,b}\|_2 + \frac{\langle s_{a,b}, c_{a,b}\rangle}{\|c_{a,b}\|_2} \ge \|c_{a,b}\|_2 - \|s_{a,b}\|_2 .$$

Taking the positive part, squaring, and then taking the supremum over a, b proves the displayed lower bound for Δ_ϕ.

For the concrete approximation bound, define

$$e_{a,b}(X) = \sum_{k=1}^{K} \pi_k(X)\eta_{k,a,b}(X), \qquad d(X) = \sum_{k=1}^{K} \{\pi_k^\star(X) - \pi_k(X)\}\nu_k(X).$$

Then $s_{a,b} = e_{a,b} + d$, so $(e_{a,b} + d)^2 \le 2e_{a,b}^2 + 2d^2$. If $|\nu_k(X)| \le M_\nu$, then

$$|d(X)| \le M_\nu \sum_{k=1}^{K} |\pi_k^\star(X) - \pi_k(X)| = M_\nu \left\|\pi^\star(X) - \pi(X)\right\|_1 .$$

Finally, since $\pi(X)$ is a probability vector, Jensen's inequality gives

$$\left(\sum_{k=1}^{K} \pi_k(X)\eta_{k,a,b}(X)\right)^2 \le \sum_{k=1}^{K} \pi_k(X)\eta_{k,a,b}(X)^2.$$

Combining these displays proves the last bound. □

Proposition 9 (Expert calibration curvature)**.** *Suppose $m(x) = \sum_k \pi_k^\star(x)\rho_k(\mu_k(x))$ and define $\delta_k(t) = \rho_k(t) - t$. For each k, let I_k be an interval containing the support of $\mu_k(X)$. Assume each δ_k is twice differentiable on I_k, $\sup_{t\in I_k} |\delta_k''(t)| \le H_k$, and $|\rho_k(\mu_k(X))| \le M_\rho$ almost surely. Then*

$$\mathcal{A}_\phi \le \frac{1}{2} \inf_{c_1\in I_1,\dots,c_K\in I_K} \sum_{k=1}^{K} H_k^2 \mathbb{E}[\pi_k(X)\{\mu_k(X) - c_k\}^4] + 2M_\rho^2 E_{\text{gate}},$$

where $E_{\text{gate}} := \mathbb{E}\left\|\pi^\star(X) - \pi(X)\right\|_1^2$. Consequently, if the realized gate is correct and, for every k, $\delta_k(\mu_k(X))$ is affine in $\mu_k(X)$ almost surely, then $\mathcal{A}_\phi = 0$.

Proof. Under the calibrated-expert representation,

$$g(x) = \sum_{k=1}^{K} \pi_k(x)\delta_k(\mu_k(x)) + \sum_{k=1}^{K} \{\pi_k^\star(x) - \pi_k(x)\}\rho_k(\mu_k(x)).$$

Fix anchors $c_k \in I_k$. For each expert, Taylor's theorem on I_k gives, for $t \in I_k$,

$$\delta_k(t) = A_k + B_k t + r_{k,2}(t), \qquad A_k = \delta_k(c_k) - c_k\delta_k'(c_k), \qquad B_k = \delta_k'(c_k),$$

with

$$|r_{k,2}(t)| \le \frac{H_k}{2}|t - c_k|^2$$

on the support of $\mu_k(X)$. Define

$$C_1(x) = \sum_{k=1}^{K} \pi_k(x)\{A_k + B_k\mu_k(x)\},$$

$$B_2(x) = \sum_{k=1}^{K} \pi_k(x) r_{k,2}(\mu_k(x)), \qquad G_{\text{gate}}(x) = \sum_{k=1}^{K} \{\pi_k^\star(x) - \pi_k(x)\}\rho_k(\mu_k(x)).$$

Then $g = C_1 + B_2 + G_{\text{gate}}$. The first-order term C_1 belongs to $\mathcal{H}_\phi$ because

$$\sum_{k=1}^{K} A_k\pi_k(x) = A_K + \sum_{k=1}^{K-1} (A_k - A_K)\pi_k(x),$$

and $\sum_k B_k\pi_k(x)\mu_k(x)$ is a linear combination of the product coordinates. Thus,

$$\mathcal{A}_\phi \le \mathbb{E}[(g(X) - C_1(X))^2] = \mathbb{E}[(B_2(X) + G_{\text{gate}}(X))^2] \le 2\mathbb{E}[B_2(X)^2] + 2\mathbb{E}[G_{\text{gate}}(X)^2].$$

For the curvature term, Jensen's inequality with the probability weights $\pi_k(X)$ gives

$$\begin{aligned} B_2(X)^2 &= \left(\sum_{k=1}^{K} \pi_k(X) r_{k,2}(\mu_k(X)) \right)^2 \\ &\le \sum_{k=1}^{K} \pi_k(X) r_{k,2}(\mu_k(X))^2 \\ &\le \frac{1}{4} \sum_{k=1}^{K} \pi_k(X) H_k^2 \{\mu_k(X) - c_k\}^4. \end{aligned}$$

For the gate term, the boundedness assumption gives

$$|G_{\mathrm{gate}}(X)| \le M_\rho \sum_{k=1}^{K} |\pi_k^\star(X) - \pi_k(X)| = M_\rho \left\| \pi^\star(X) - \pi(X) \right\|_1 .$$

Substituting the two bounds into the preceding display proves

$$\mathcal{A}_\phi \le \frac{1}{2} \sum_{k=1}^{K} H_k^2 \mathbb{E}[\pi_k(X)\{\mu_k(X) - c_k\}^4] + 2M_\rho^2 \mathbb{E} \left\| \pi^\star(X) - \pi(X) \right\|_1^2 .$$

Taking the infimum over anchors $c_k \in I_k$ gives the stated bound. If $\pi^\star = \pi$ and $\delta_k(\mu_k(X)) = A_k + B_k \mu_k(X)$ almost surely for every k, then $g(X) = \sum_k \pi_k(X)\{A_k + B_k \mu_k(X)\} \in \mathcal{H}_\phi$, so $\mathcal{A}_\phi = 0$. □

A.3 First-Order Readout Interpretation

The preceding decompositions use a routed reference model to interpret the residual structure. A complementary view only uses the realized readout Z and treats the residual regression function as an unknown function on the teacher-feature space.

Proposition 10 (First-order readout approximation)**.** *Let $h(z) = \mathbb{E}[g(X) \mid Z(X) = z]$ and write*

$$g(X) = h(Z(X)) + \xi(X), \qquad \mathbb{E}[\xi(X) \mid Z(X)] = 0.$$

If h is twice differentiable on a convex set containing the support of Z and the anchor z_0, and its Hessian has operator norm at most H on that set, then

$$\mathcal{A}_\phi \le \mathbb{E}[\xi(X)^2] + \frac{H^2}{4} \mathbb{E} \left\| Z(X) - z_0 \right\|_2^4 .$$

Proof. Since $\mathcal{H}_\phi$ contains affine functions of Z, define

$$a_{z_0}(z) = h(z_0) + \nabla h(z_0)^\top (z - z_0).$$

Then

$$\mathcal{A}_\phi \le \mathbb{E}[(g(X) - a_{z_0}(Z(X)))^2].$$

The decomposition $g = h + \xi$ and the conditional mean-zero property of ξ give

$$\mathbb{E}[(g(X) - a_{z_0}(Z(X)))^2] = \mathbb{E}[(h(Z(X)) - a_{z_0}(Z(X)))^2] + \mathbb{E}[\xi(X)^2].$$

Taylor's theorem gives

$$|h(Z) - a_{z_0}(Z)| \le \frac{H}{2} \left\| Z - z_0 \right\|_2^2 ,$$

which proves the claim after squaring and taking expectations. □

A.4 kNN Residual Imputation

The finite-sample theorem uses

$$h(z) = \mathbb{E}[Y - \hat{f}(X) \mid Z = z] = \mathbb{E}[g(X) \mid Z = z].$$

Since $\psi(X)$ is measurable with respect to Z, the projection onto $\mathcal{H}_\phi$ can be written in terms of h.

Lemma 11 (Equivalent projection through $h(Z)$)**.** *Let $h(z) = \mathbb{E}[g(X) \mid Z = z]$. Then*

$$\arg\min_{\theta \in \mathbb{R}^{2K}} \mathbb{E}[(g(X) - \psi(X)^\top \theta)^2] = \arg\min_{\theta \in \mathbb{R}^{2K}} \mathbb{E}[(h(Z) - \psi(X)^\top \theta)^2].$$

Moreover, for $U(Z) = h(Z) - \psi(X)^\top \theta^\star$ with $\theta^\star$ denoting this common minimizer,

$$\mathbb{E}[\psi(X)U(Z)] = 0.$$

Proof. For any θ, since $\psi(X)$ is measurable with respect to Z,

$$g(X) - \psi(X)^\top \theta = \{g(X) - h(Z)\} + \{h(Z) - \psi(X)^\top \theta\}.$$

The cross term has expectation zero because $\mathbb{E}[g(X) - h(Z) \mid Z] = 0$ and $h(Z) - \psi(X)^\top \theta$ is Z-measurable. Hence

$$\mathbb{E}[(g(X) - \psi(X)^\top \theta)^2] = \mathbb{E}[(h(Z) - \psi(X)^\top \theta)^2] + \mathbb{E}[(g(X) - h(Z))^2].$$

The last term is independent of θ, so the two argmin sets coincide. The orthogonality identity follows from the first-order optimality condition for the second objective. □

For a query point z, let $N_{k_{\mathrm{nn}}}(z)$ be the indices of the k_{nn} nearest labeled feature vectors among $Z_1^s, \ldots, Z_{n_s}^s$. Let

$$\widehat{h}_{\mathrm{nn}}(z) = \sum_{i \in N_{k_{\mathrm{nn}}}(z)} a_i(z) R_i^s$$

where the weights satisfy $a_i(z) \geq 0$, $\sum_{i \in N_{k_{\mathrm{nn}}}(z)} a_i(z) = 1$, and $a_i(z) = 0$ outside $N_{k_{\mathrm{nn}}}(z)$. The inverse-distance rule used by the algorithm is one example; the following bounds only require normalized nonnegative weights.

Lemma 12 (kNN imputation bias–variance)**.** *Assume h is s-Hölder on the support of Z: for some $s \in (0, 1]$ and $L_h < \infty$,*

$$|h(z) - h(z')| \leq L_h \left\| z - z' \right\|_2^s.$$

Assume also that $\zeta = Y - \hat{f}(X) - h(Z)$ satisfies $\mathbb{E}[\zeta \mid Z] = 0$ and $\mathrm{Var}(\zeta \mid Z) \leq \sigma_\zeta^2$ almost surely. Let

$$\rho_{k_{\mathrm{nn}}}(z) = \max_{i \in N_{k_{\mathrm{nn}}}(z)} \left\| Z_i^s - z \right\|_2, \qquad \nu_{k_{\mathrm{nn}}}(z) = \sum_{i \in N_{k_{\mathrm{nn}}}(z)} a_i(z)^2.$$

Conditionally on the labeled covariates,

$$\mathbb{E}\{(\widehat{h}_{\mathrm{nn}}(z) - h(z))^2 \mid Z_1^s, \ldots, Z_{n_s}^s\} \leq L_h^2 \rho_{k_{\mathrm{nn}}}(z)^{2s} + \sigma_\zeta^2 \nu_{k_{\mathrm{nn}}}(z).$$

Consequently, for an independent query Z,

$$\mathbb{E}[(\widehat{h}_{\mathrm{nn}}(Z) - h(Z))^2] \leq L_h^2 \mathbb{E}[\rho_{k_{\mathrm{nn}}}(Z)^{2s}] + \sigma_\zeta^2 \mathbb{E}[\nu_{k_{\mathrm{nn}}}(Z)].$$

Proof. Write $R_i^s = h(Z_i^s) + \zeta_i^s$. Conditional on the labeled covariates,

$$\widehat{h}_{\mathrm{nn}}(z) - h(z) = \sum_{i \in N_{k_{\mathrm{nn}}}(z)} a_i(z)\{h(Z_i^s) - h(z)\} + \sum_{i \in N_{k_{\mathrm{nn}}}(z)} a_i(z)\zeta_i^s.$$

The first term has magnitude at most $L_h \rho_{k_{\mathrm{nn}}}(z)^s$. The second term has conditional mean zero and conditional variance at most $\sigma_\zeta^2 \sum_i a_i(z)^2$. The cross term vanishes after conditioning on the labeled covariates. Averaging over the independent query proves the result. □

Lemma 13 (kNN rate under affine teacher-feature mass dimension)**.** *Assume the conditions of Lemma 12. Assume further that the distribution of Z has mass dimension d_ϕ: there exist constants $c_-, c_+, r_0 > 0$ such that, for every support point z and every $0 < r \le r_0$,*

$$c_- r^{d_\phi} \le \mathbb{P}(\|Z - z\|_2 \le r) \le c_+ r^{d_\phi}. \tag{7}$$

Assume also $\|Z\|_2 \le L_Z$ almost surely. If $1 \le k_{\mathrm{nn}} \le n_s/2$, then

$$\mathbb{E}[\rho_{k_{\mathrm{nn}}}(Z)^{2s}] \le C_{\mathrm{geo}} \left(\frac{k_{\mathrm{nn}}}{n_s}\right)^{2s/d_\phi},$$

where C_{geo} depends only on $(s, d_\phi, c_-, c_+, r_0, L_Z)$. Hence

$$\mathbb{E}[(\widehat{h}_{\mathrm{nn}}(Z) - h(Z))^2] \le C_{\mathrm{geo}} L_h^2 \left(\frac{k_{\mathrm{nn}}}{n_s}\right)^{2s/d_\phi} + \sigma_\zeta^2 \mathbb{E}[\nu_{k_{\mathrm{nn}}}(Z)].$$

If the weights are balanced in the sense that

$$\mathbb{E}[\nu_{k_{\mathrm{nn}}}(Z)] \le \frac{C_\nu}{k_{\mathrm{nn}}}, \tag{8}$$

then

$$\mathbb{E}[(\widehat{h}_{\mathrm{nn}}(Z) - h(Z))^2] \le C_{\mathrm{geo}} L_h^2 \left(\frac{k_{\mathrm{nn}}}{n_s}\right)^{2s/d_\phi} + \frac{\sigma_\zeta^2 C_\nu}{k_{\mathrm{nn}}}.$$

Choosing $k_{\mathrm{nn}} \asymp n_s^{2s/(2s+d_\phi)}$ gives the rate $n_s^{-2s/(2s+d_\phi)}$.

Proof. For a fixed query z, let $B(z, r) = \{z' : \|z' - z\|_2 \le r\}$ and $p_z(r) = \mathbb{P}\{Z \in B(z, r)\}$. If $p_z(r) \ge 2k_{\mathrm{nn}}/n_s$, a binomial Chernoff bound gives

$$\mathbb{P}\{\rho_{k_{\mathrm{nn}}}(z) > r \mid Z = z\} = \mathbb{P}\left\{\sum_{i=1}^{n_s} \mathbf{1}\{Z_i^s \in B(z, r)\} < k_{\mathrm{nn}}\right\} \le \exp\{-c n_s p_z(r)\}$$

for a universal constant $c > 0$. Let $r_\star = (2k_{\mathrm{nn}}/(c_- n_s))^{1/d_\phi}$. When $r_\star \le r_0$, the lower-mass condition implies $p_z(r) \ge 2k_{\mathrm{nn}}/n_s$ for all $r \in [r_\star, r_0]$, so

$$\mathbb{P}\{\rho_{k_{\mathrm{nn}}}(z) > r \mid Z = z\} \le \exp\{-c n_s r^{d_\phi}\}, \qquad r \in [r_\star, r_0].$$

Since $\|Z\|_2 \le L_Z$, $\rho_{k_{\mathrm{nn}}}(z) \le 2L_Z$. Integrating the tail probability gives

$$\mathbb{E}[\rho_{k_{\mathrm{nn}}}(z)^{2s}] = \int_0^{2L_Z} 2s\, r^{2s-1} \mathbb{P}\{\rho_{k_{\mathrm{nn}}}(z) > r \mid Z = z\}\, dr \le C_{\mathrm{geo}} \left(\frac{k_{\mathrm{nn}}}{n_s}\right)^{2s/d_\phi}.$$

If $r_\star > r_0$, then k_{nn}/n_s is bounded below by a constant depending only on c_- and r_0, and the deterministic bound $\rho_{k_{\mathrm{nn}}}(z) \le 2L_Z$ is absorbed by increasing C_{geo}. The bound is uniform in z, hence also holds after averaging over Z. Combining this bound with Lemma 12 proves the first displays. Balancing $(k_{\mathrm{nn}}/n_s)^{2s/d_\phi}$ and $1/k_{\mathrm{nn}}$ gives the stated choice of k_{nn}. □

Remark 5 (Balanced weights). Uniform kNN weights satisfy $\nu_{k_{\mathrm{nn}}}(Z) = 1/k_{\mathrm{nn}}$ exactly. Inverse-distance weights satisfy (8) when the distances inside local kNN balls are comparable or when the stabilizer keeps one neighbor from receiving nearly all the mass. Without (8), Lemma 13 remains valid with the exact quantity $\mathbb{E}[\nu_{k_{\mathrm{nn}}}(Z)] \le 1$, but the imputation variance term need not improve as $1/k_{\mathrm{nn}}$.

Corollary 14 (High-probability imputation event)**.** *Assume $n_u \ge 1$. Let $\delta_j = \widehat{h}_{\mathrm{nn}}(Z_j^u) - h(Z_j^u)$. Suppose that, conditionally on the labeled sample,*

$$\mathbb{E}[(\widehat{h}_{\mathrm{nn}}(Z) - h(Z))^2] \le \rho^2$$

for an independent $Z \sim P_Z$, and suppose $|\widehat{h}_{\mathrm{nn}}(Z) - h(Z)| \le M_{\mathrm{imp}}$ almost surely. Then, conditionally on the labeled sample, with probability at least $1 - \delta$ over the unlabeled covariates,

$$\frac{1}{n_u} \sum_{j=1}^{n_u} \delta_j^2 \le \rho^2 + C M_{\mathrm{imp}}^2 \sqrt{\frac{\log(1/\delta)}{n_u}}$$

for a universal constant C. If the squared imputation errors are sub-exponential, the same display holds with the usual Bernstein replacement.

Moreover, if Lemma 13 yields the unconditional bound

$$\bar{\rho}^2_{k_{\mathrm{nn}}} = C_{\mathrm{geo}} L_h^2 \left(\frac{k_{\mathrm{nn}}}{n_s}\right)^{2s/d_\phi} + \frac{\sigma_\zeta^2 C_\nu}{k_{\mathrm{nn}}},$$

then for any $\eta, \delta \in (0,1)$, with probability at least $1-\eta-\delta$ over the labeled sample and unlabeled covariates,

$$\frac{1}{n_u}\sum_{j=1}^{n_u} \delta_j^2 \le \frac{\bar{\rho}^2_{k_{\mathrm{nn}}}}{\eta} + C M_{\mathrm{imp}}^2 \sqrt{\frac{\log(1/\delta)}{n_u}}.$$

For fixed η, the rate in n_s is unchanged.

Proof. Conditional on the labeled sample, $\delta_1^2, \dots, \delta_{n_u}^2$ are i.i.d., bounded by M_{imp}^2, and have mean at most ρ^2. Hoeffding's inequality gives the first display, and the sub-exponential version follows from Bernstein's inequality.

For the unconditional statement, let

$$\rho^2(\mathcal{S}) = \mathbb{E}[(\widehat{h}_{\mathrm{nn}}(Z) - h(Z))^2 \mid \mathcal{S}],$$

where the expectation is over an independent query and the imputer is built from the labeled sample. Lemma 13 gives $\mathbb{E}_{\mathcal{S}}\rho^2(\mathcal{S}) \le \bar{\rho}^2_{k_{\mathrm{nn}}}$. Markov's inequality gives $\rho^2(\mathcal{S}) \le \bar{\rho}^2_{k_{\mathrm{nn}}}/\eta$ with probability at least $1-\eta$ over the labeled sample. Applying the conditional bound on this event and taking a union bound proves the claim. □

A.5 Weighted Ridge Perturbation

Let

$$W = n_s + w_u n_u, \qquad \lambda = \alpha/W, \qquad N_w = \frac{(n_s + w_u n_u)^2}{n_s + w_u^2 n_u}, \qquad \tau_u = \frac{w_u n_u}{W},$$

and let $\mathsf{P} = \mathrm{diag}(0, 1, \dots, 1)$ so that the ridge penalty leaves the intercept unpenalized. For the kNN step, write $R = Y - \hat{f}(X)$ and $h(z) = \mathbb{E}[R \mid Z = z]$. Let $\widehat{h}_{\mathrm{nn}}(Z_j^u)$ be the inverse-distance kNN imputation of $h(Z_j^u)$ from labeled residuals, and let $\mathcal{E}_{\mathrm{imp}}(r)$ denote the event

$$\frac{1}{n_u}\sum_{j=1}^{n_u} \{\widehat{h}_{\mathrm{nn}}(Z_j^u) - h(Z_j^u)\}^2 \le r^2.$$

Remark 6. When $n_u = 0$, the same argument gives the labeled-only ATF ridge bound by setting $W = N_w = n_s$, $\tau_u = 0$, and omitting the event $\mathcal{E}_{\mathrm{imp}}(r)$.

Assumption 2 (Residual noise and projection residual)**.** The residual noise $\zeta = R - h(Z)$ is conditionally sub-Gaussian with parameter σ_ζ given Z, and $|h(Z) - \psi(X)^\top \theta^\star| \le B_\star$ almost surely, where $\theta^\star$ is the oracle ATF projection coefficient.

We prove Theorem 2. More precisely, under Assumptions 1–2, if the sample-size conditions in Proposition 3 hold, then with probability at least $1 - \delta - \mathbb{P}(\mathcal{E}_{\mathrm{imp}}(r)^c)$, the theorem holds with the following admissible explicit choice for the finite-sample term:

$$\varepsilon_{\mathrm{stat}} = C\frac{L^2}{\kappa^2}\left[L^2 B_\star^2 \frac{(2K)\log(16K/\delta)}{N_w} + L^2\sigma_\zeta^2 \frac{(2K)\log(16K/\delta) n_s}{W^2} + L^2\tau_u^2 r^2 + \lambda^2 \left\|\mathsf{P}\theta^\star\right\|_2^2\right],$$

with a universal constant C. Define the weighted empirical Gram matrix and right-hand side

$$\widehat{\Sigma}_w = \frac{1}{W}\left\{\sum_{i=1}^{n_s} \psi_i^s \psi_i^{s\top} + w_u \sum_{j=1}^{n_u} \psi_j^u \psi_j^{u\top}\right\},$$

$$\widehat{b}_w = \frac{1}{W}\left\{\sum_{i=1}^{n_s}\psi_i^s R_i^s + w_u\sum_{j=1}^{n_u}\psi_j^u \widehat{h}_{\mathrm{nn}}(Z_j^u)\right\}.$$

The ridge normal equation is

$$(\widehat{\Sigma}_w + \lambda \mathsf{P})\widehat{\theta} = \widehat{b}_w. \tag{9}$$

Lemma 15 (Weighted covariance concentration)**.** *Under Assumption 1 and $w_u \in [0,1]$, there is a universal constant $c > 0$ such that, for any $\delta \in (0,1)$, if*

$$W \geq c\,\frac{L^2}{\kappa}\log\frac{16K}{\delta} \qquad \textit{and} \qquad N_w \geq c\,\frac{L^2\,\|\Sigma\|_{\mathrm{op}}}{\kappa^2}\log\frac{16K}{\delta},$$

then

$$\lambda_{\min}(\widehat{\Sigma}_w) \geq \frac{\kappa}{2}$$

with probability at least $1-\delta/4$.

Proof. Write $\widehat{\Sigma}_w - \Sigma = \sum_t X_t$ with $X_t = (c_t/W)(\psi_t\psi_t^\top - \Sigma)$, where t runs over the labeled and unlabeled rows, $c_t = 1$ on labeled rows, and $c_t = w_u$ on unlabeled rows. The X_t are independent, mean-zero, and self-adjoint. Since $\psi_t\psi_t^\top$ and Σ are positive semidefinite with operator norms at most L^2, $\|X_t\|_{\mathrm{op}} \leq c_t L^2/W \leq L^2/W$. For the variance,

$$\mathbb{E}\big[(\psi\psi^\top - \Sigma)^2\big] \preceq \mathbb{E}\big[(\psi\psi^\top)^2\big] = \mathbb{E}\big[\|\psi\|_2^2\,\psi\psi^\top\big] \preceq L^2\Sigma,$$

so

$$\Big\|\sum_t \mathbb{E}[X_t^2]\Big\|_{\mathrm{op}} \leq \frac{\sum_t c_t^2}{W^2}\,L^2\,\|\Sigma\|_{\mathrm{op}} = \frac{L^2\,\|\Sigma\|_{\mathrm{op}}}{N_w},$$

using $\sum_t c_t^2 = n_s + w_u^2 n_u$. The matrix Bernstein inequality (Tropp, 2012) in dimension $2K$ gives

$$\mathbb{P}\{\big\|\widehat{\Sigma}_w - \Sigma\big\|_{\mathrm{op}} \geq \kappa/2\} \leq 4K\exp\left\{-\frac{\kappa^2/8}{L^2\,\|\Sigma\|_{\mathrm{op}}/N_w + L^2\kappa/(6W)}\right\}.$$

Under the two stated conditions with $c = 16$, each term in the denominator is at most $\kappa^2/\{16\log(16K/\delta)\}$, so the exponent is at least $\log(16K/\delta)$ and the probability is at most $\delta/4$. Weyl's inequality then gives $\lambda_{\min}(\widehat{\Sigma}_w) \geq \lambda_{\min}(\Sigma) - \kappa/2 \geq \kappa/2$. □

Remark 7 (Dependence on the number of experts). Because $2K\kappa \leq \operatorname{tr}\Sigma = \mathbb{E}\,\|\psi\|_2^2 \leq L^2$, the factor L^2/κ is at least $2K$. If the scaled coordinates satisfy $|\psi_\ell(X)| \leq B$, then $L^2 \leq 2KB^2$, and the second condition of Lemma 15 reads $N_w \gtrsim 2K\,(B^2/\kappa)(\|\Sigma\|_{\mathrm{op}}/\kappa)\log(16K/\delta)$: linear in the number of coefficients up to conditioning and logarithmic factors. The constants and the polynomial dependence on K in the explicit $\varepsilon_{\mathrm{stat}}$ below are not optimized.

Proof of Proposition 3. The high-probability covariance statement is Lemma 15. If $w_u = 0$, then $W = n_s$ and $N_w = n_s$, so the conditions reduce to labeled-sample lower bounds. □

Lemma 16 (Empirical score concentration)**.** *Under Assumptions 1–2, with probability at least $1-\delta/2$,*

$$\left\|\frac{1}{W}\left\{\sum_{i=1}^{n_s}\psi_i^s U(Z_i^s) + w_u\sum_{j=1}^{n_u}\psi_j^u U(Z_j^u)\right\}\right\|_2^2 \leq CL^2B_\star^2\,\frac{(2K)\log(16K/\delta)}{N_w}, \tag{10}$$

where $U(Z) = h(Z) - \psi(X)^\top\theta^\star$, and

$$\left\|\frac{1}{W}\sum_{i=1}^{n_s}\psi_i^s\zeta_i^s\right\|_2^2 \leq CL^2\sigma_\zeta^2\,\frac{(2K)\log(16K/\delta)n_s}{W^2}. \tag{11}$$

Proof. By Lemma 11, $\mathbb{E}[\psi(X)U(Z)] = 0$. Each coordinate of the weighted average in (10) is a sum of independent, bounded, mean-zero random variables with absolute value bounded by $LB_\star/W$ on labeled rows and $w_u LB_\star/W$ on unlabeled rows. Hoeffding's inequality and a union bound over the $2K$ coordinates give (10); the Euclidean norm conversion contributes the factor $2K$.

For (11), condition on the covariates. Each coordinate is sub-Gaussian with proxy at most $L^2\sigma_\zeta^2 n_s/W^2$. A coordinate union bound and the same Euclidean-norm conversion give the result. □

Proof of Theorem 2. Let

$$\delta_j = \widehat{h}_{\mathrm{nn}}(Z_j^u) - h(Z_j^u), \qquad U(Z) = h(Z) - \psi(X)^\top \theta^\star.$$

Subtract

$$(\widehat{\Sigma}_w + \lambda \mathsf{P})\theta^\star = \widehat{\Sigma}_w \theta^\star + \lambda \mathsf{P}\theta^\star$$

from the normal equation (9). Using $R_i^s = h(Z_i^s) + \zeta_i^s$, we obtain

$$(\widehat{\Sigma}_w + \lambda \mathsf{P})(\widehat{\theta} - \theta^\star) = V_0 + V_1 + V_2 - \lambda \mathsf{P}\theta^\star, \tag{12}$$

where

$$V_0 = \frac{1}{W}\left\{ \sum_{i=1}^{n_s} \psi_i^s U(Z_i^s) + w_u \sum_{j=1}^{n_u} \psi_j^u U(Z_j^u) \right\},$$

$$V_1 = \frac{1}{W} \sum_{i=1}^{n_s} \psi_i^s \zeta_i^s,$$

$$V_2 = \frac{w_u}{W} \sum_{j=1}^{n_u} \psi_j^u \delta_j.$$

On the covariance event of Lemma 15,

$$\lambda_{\min}(\widehat{\Sigma}_w + \lambda \mathsf{P}) \geq \kappa/2$$

because $\lambda \mathsf{P} \succeq 0$. Therefore

$$\left\| \widehat{\theta} - \theta^\star \right\|_2 \leq \frac{2}{\kappa} \left\| V_0 + V_1 + V_2 - \lambda \mathsf{P}\theta^\star \right\|_2.$$

The projection normal equation gives

$$\mathcal{R}(\widehat{m}_{\mathrm{ATF}}) - \mathcal{A}_\phi = (\widehat{\theta} - \theta^\star)^\top \Sigma (\widehat{\theta} - \theta^\star) \leq L^2 \left\| \widehat{\theta} - \theta^\star \right\|_2^2,$$

where $\lambda_{\max}(\Sigma) \leq \mathbb{E}[\|\psi(X)\|_2^2] \leq L^2$.

Lemma 16 controls V_0 and V_1. On $\mathcal{E}_{\mathrm{imp}}(r)$, Cauchy's inequality and $\left\|\psi_j^u\right\|_2 \leq L$ give

$$\|V_2\|_2 \leq \frac{w_u}{W} \sum_{j=1}^{n_u} \left\|\psi_j^u\right\|_2 |\delta_j| \leq L \frac{w_u n_u}{W} \left(\frac{1}{n_u} \sum_{j=1}^{n_u} \delta_j^2 \right)^{1/2} \leq L\tau_u r.$$

Finally,

$$\|V_0 + V_1 + V_2 - \lambda \mathsf{P}\theta^\star\|_2^2 \leq 4\left(\|V_0\|_2^2 + \|V_1\|_2^2 + \|V_2\|_2^2 + \lambda^2 \|\mathsf{P}\theta^\star\|_2^2 \right).$$

Combining these displays and absorbing numerical constants into C proves the risk bound. The probability follows by a union bound over the covariance, score, and imputation events.

The improvement certificates follow from the projection identities in Theorem 1:

$$\mathcal{R}(\hat{f}) = \mathcal{A}_\phi + \Delta_\phi, \qquad \mathcal{R}(m_{\mathrm{sc}}) = \mathcal{A}_\phi + G_{\phi|\mathrm{sc}}.$$

Subtracting the risk bound from these identities gives the two lower bounds. □

B PROTOCOLS FOR THE NATURAL-SHIFT, SYNTHETIC-SUITE, AND BIKE-SHARING EXPERIMENTS

Shared implementation. All teachers are trained on source data only and then frozen. ATF-KNN and ATF-NO-KNN follow Section 3.2: the features (3) are standardized with statistics of $\mathcal{S}_X \cup \mathcal{U}$, the ridge fit (5) has an unpenalized intercept ($2K$ coefficients), and the imputation (4) uses $\epsilon = 10^{-10}$ and at most as many neighbors as labeled points. Hyperparameters are selected by 3-fold cross-validation (CV) on the labeled set, with ATF-KNN imputing residuals from the training folds only. Unless stated otherwise, $\alpha \in \{0.25, 0.5, 1, 1.5, 2, 4\}$ and $w_u = 0.5$; MoE-adapted PP-SSL and SADA use the design $[1, \tilde{\phi}(x)]$ (Appendix F), PP-SSL is tuned over learning rate $\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}\} \times$ steps $\{200, 800\} \times$ initial weight $\{0, 0.5\} \times$ initial ridge $\{0.1, 1\}$, and SADA uses ridge 0.2 and shrinkage 0.1. The Electricity, ACSIncome, and synthetic teachers are trained with AdamW (learning rate 10^{-3}, weight decay 10^{-4}), early stopping on source validation data, and the load-balancing penalty $\rho K \sum_k (\bar{\pi}_k - 1/K)^2$, with $\bar{\pi}_k$ the minibatch mean of π_k. The white-box controls apply L2-SP fine-tuning (Li et al., 2018) to a copy of the teacher: from the teacher weights θ_0, they update the output layer of every expert (*expert head*), the gate (*router*), or all parameters (*full*) by full-batch AdamW without weight decay on the target loss plus $\lambda \|\theta - \theta_0\|_2^2$ over the updated parameters θ. Unless stated otherwise, the learning rate $\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}\}$ and $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ are selected by labeled CV with early stopping on the held-out fold, and the model is refit from θ_0 on all labeled examples for the rounded median of the fold-wise best epoch counts (zero epochs returns the teacher). Fine-tuning uses no unlabeled data, and no method uses unlabeled or evaluation outcomes.

B.1 ELECTRICITY: NATURAL TEMPORAL SHIFT

Data and teacher. We use the hourly version of the UCI Electricity Load Diagrams 2011–2014 data (Trindade, 2015), restricted to a fixed-seed sample of 128 clients among those with nonzero load in at least 95% of the 2012 hours. An example is a client-hour with outcome $\log(1 + \text{consumption})$ and 16 covariates: eight principal-component scores of the client's 2012 mean log-load profile over the 168 hours of the week, a linear time trend, sine–cosine encodings of hour, weekday, and day of year, and a weekend indicator. The teacher, a top-2 MoE with $K = 4$ experts (MLPs of width 32), is trained on 120,000 client-hours from 2012 with $\rho = 0.05$, early-stopped on 40,000 held-out client-hours from 2013, and frozen. Profiles, components, and standardization use 2012 data only; no lagged outcome is used, and no 2014 data enter client selection, preprocessing, or teacher construction.

Samples and methods. Three disjoint random sets of 2014 client-hours are drawn once: a 100,000-example adaptation pool, an evaluation set ($n_e = 60{,}000$), and a 20,000-example diagnostic set. Each of 30 paired repetitions permutes the pool and takes $n_s = 40$ labeled and $n_u = 6{,}000$ unlabeled examples (labels discarded); the metric is RMSE. The methods (fitted coefficients in parentheses) are the teacher (0); mean correction (1); scalar affine correction, a ridge fit of y on $[1, \hat{f}]$ with $\alpha \in \{0, 10^{-4}, 10^{-2}, 0.1, 1, 10\}$ (2); target-only ridge on the covariates with $\alpha \in \{0.1, 1, 10, 100\}$ (17); PP-SSL and SADA (8 each); ATF-NO-KNN and ATF-KNN with $k_{\text{nn}} = 15$ (8 each); and fine-tuning of the expert heads (132) and of the router (1,732) for at most 120 epochs (patience 15).

Routing diagnostics (computed once; used by no method). (i) *Shuffled routing*: on the 2013 validation and 2014 diagnostic sets, the RMSE of $\sum_k \pi_k \mu_k$ is compared with that of 200 predictions whose routing vectors are permuted across examples, and we count the experts that are the most probable route for at least one example. (ii) *Routing-region residual test*: the between-region sum of squares of the teacher residuals on the diagnostic set, with regions given by the most probable expert, is compared with 1,000 random permutations of the residuals, $p = (1 + \#\{\text{permuted} \geq \text{observed}\})/1001$. (iii) *Cross-fitted comparison*: a scalar affine ridge on $[1, \hat{f}]$ and a ridge of $y - \hat{f}$ on $[1, \tilde{\phi}]$, with penalties chosen by inner 5-fold CV, are compared by 5-fold out-of-fold RMSE on the diagnostic set; this uses all 20,000 diagnostic labels and is not a scarce-label method.

B.2 ACSINCOME: CALIFORNIA TO MICHIGAN

Data and teacher. We use the Folktables ACSIncome task (Ding et al., 2021) on the 2018 ACS 1-Year person files, with its standard sample filter, ten standard features, and binary label of income

above $50,000; features are standardized or one-hot encoded using California training data only (815 inputs). California (195,665 persons) is split 80/20 into training and validation data, and Michigan (50,008 persons) 50/50 into an adaptation pool and a fixed evaluation set of 25,004 examples each, both stratified by label. The teacher has $K = 4$ experts (fixed in advance); gate and experts are MLPs of width 64 with softmax and sigmoid outputs, so that $\hat{p}(x) = \sum_k \pi_k(x) q_k(x)$. It is trained on California by binary cross-entropy, early-stopped on the California validation log loss, which also selects $\rho \in \{0, 0.01, 0.05\}$ ($\rho = 0.05$), and frozen; no Michigan data are used.

Samples, correction, and methods. Each of 50 paired repetitions draws from the Michigan pool nested, class-stratified labeled sets of sizes $n_s \in \{10, 20, 30, 40\}$ and $n_u = 10{,}000$ unlabeled examples from the rest of the pool, shared across budgets; all methods are evaluated on the same $n_e = 25{,}004$ examples. Corrections act in probability space: residuals are $y - \hat{p}(x)$, $\phi(x) = [\pi_{1:K-1}(x), \pi(x) \odot q(x)]$, and corrected probabilities are clipped to $[10^{-6}, 1 - 10^{-6}]$. Metrics are the Brier score (primary) and accuracy (threshold 0.5). The methods are the teacher; scalar residual ridge of $y - \hat{p}$ on $[1, \hat{p}]$ (2 coefficients); target-only ℓ_2-regularized logistic regression (816); PP-SSL and SADA (8 each); a same-readout RBF kernel ridge regression of the labeled residuals on $\tilde{\phi}$ (n_s dual coefficients); fine-tuning of the expert heads (260), the router (56,644), and the full teacher (265,800) with binary cross-entropy for at most 200 epochs (patience 20); and ATF-NO-KNN and ATF-KNN (8 each). Selection uses stratified CV on the Brier score over $\alpha \in \{0.01, 0.1, 1, 10, 100\}$ for the scalar and ATF ridges, jointly with $k_{\text{nn}} \in \{3, 5, 10, 15\}$ for ATF-KNN; $C \in \{0.01, 0.1, 1, 10\}$ for the target-only logistic regression; SADA ridge $\{0.01, 0.1, 1, 10\}$ $\times$ shrinkage $\{0, 0.1, 0.25, 0.5\}$; and RBF penalty $\{0.01, 0.1, 1, 10, 100\}$ $\times$ bandwidth multiplier $\{0.5, 1, 2\}$ of the median heuristic. Adaptation time (Table 10) is the CPU wall-clock time per method, repetition, and budget for CV, the final fit, and evaluation-set prediction; the shared teacher forward pass is excluded for frozen-interface methods, whereas fine-tuning times include all passes of the adapted network.

B.3 TEACHER-INDEPENDENT SYNTHETIC SUITE

Worlds and target mechanisms. In each of 30 seeded worlds, a latent $u \in \mathbb{R}^4$ follows a six-component Gaussian mixture, and the 20 covariates are ten noisy random linear projections of u, four nonlinear functions of random projections, and six $\mathcal{N}(0, 1)$ nuisance coordinates. The target changes the uniform source mixture weights to $(0.30, 0.25, 0.18, 0.12, 0.09, 0.06)$ in random order and moves each component mean by a random vector of norm 0.35. The source regression function $m_S(u, z)$ is a random nonlinear function with component-specific terms, standardized under the source distribution, and outcomes add $\mathcal{N}(0, 0.5^2)$ noise. The target conditional mean is $m_T = m_S + 0.75\,\tilde{g}$, where $\tilde{g}$ is, after standardization under the target covariate distribution, either a *smooth* random combination of sine, tanh, and product terms in random projections of u, or a *rough* piecewise-constant function with $\mathcal{N}(0, 1)$ values on the eight cells cut by three random hyperplanes; neither depends on z. The *scalar* control is $m_T = 1.15\, m_S + 0.25$. All generator parameters are drawn from the world seed, and all samples and outcomes are generated before the teacher is initialized, by code that receives no teacher quantity; each target shift is thus fixed before, and independently of, the teacher.

Teacher, samples, and methods. In each world, a top-2 MoE with $K = 4$ experts, fewer than the six generator components, is trained on 20,000 standardized source examples with $\rho = 0.02$, early-stopped on 5,000 source-validation examples, and frozen. Each world has 80 labeled candidates, $n_u = 5{,}000$ unlabeled, and $n_e = 10{,}000$ evaluation examples, shared across mechanisms with independent outcome noise; the budgets $n_s \in \{10, 20, 40, 80\}$ are nested prefixes of the candidates; the metric is RMSE against the noise-free m_T. The methods are the teacher; mean correction; scalar affine correction; target-only ridge on the covariates and on 256 random Fourier features; PP-SSL with fixed settings (800 steps, learning rate $3 \cdot 10^{-4}$) and SADA; ATF-NO-KNN and ATF-KNN with $k_{\text{nn}} = 5$ (8 coefficients); and, under the smooth mechanism, fine-tuning of the router (5,764 parameters) and the expert heads (260) with learning rate 10^{-3}, $\lambda \in \{0.1, 1, 10, 100\}$, and at most 80 epochs (patience 10). Frozen-interface methods run on CPU, and fine-tuning on the teachers' GPU.

Routed-feature permutation control. With $c_k = \pi_k \mu_k$, so that $\hat{f} = \sum_k c_k$, the readout is reparameterized as $\psi = [\hat{f}, \pi_{1:K-1}, c_1 - c_K, \ldots, c_{K-1} - c_K]$, which with an intercept spans the same space as $[1, \phi]$. We compare residual ridge fits without kNN on ψ and on ψ with all coordinates

except $\hat{f}$ permuted jointly across examples within each of the labeled, unlabeled, and evaluation sets; this keeps $\hat{f}$ and all marginals but breaks the link between each example and its routed coordinates.

B.4 Bike-Sharing with white-box controls

Data (Fanaee-T & Gama, 2014), preprocessing, and the frozen five-expert neural MoE follow the configuration in Appendix E.1: the outcome is $\log(1 + \text{count})$, the metric is RMSE, covariates are standardized with a scaler fit on the covariates of all records (never on outcomes), and the teacher is pretrained on $n_{\text{pre}} = 9{,}000$ examples with regime-dependent subsampling and feature corruption. In each of 30 repetitions, a random permutation of the records yields the pretraining pool, 20 labeled, $n_u = 5{,}000$ unlabeled, $n_e = 3{,}000$ evaluation, and 20 further labeled examples; $n_s = 10$, 20, and 40 use the first 10, the first 20, and all 40 labeled examples, with teacher, unlabeled, and evaluation sets shared across budgets. The methods are the teacher; mean correction; the target-only fit (labeled-only polynomial ridge); PP-SSL and SADA; ATF-NO-kNN and ATF-kNN with $k_{\text{nn}} = 15$ (10 coefficients); and fine-tuning of the expert heads (165 parameters), the router (1,477), and the full MoE (2,922) for at most 200 epochs (patience 20). The scalar-teacher-mix baseline (Appendix F) is also run as a diagnostic control. Times are CPU wall-clock times including teacher forward passes.

B.5 Statistical protocol

Within a repetition (a world in the synthetic suite), all methods share the frozen teacher and the labeled, unlabeled, and evaluation samples, labeled sets are nested across budgets, and comparisons are paired by repetition. For Electricity, ATF-kNN is compared with each of the other nine methods by two-sided Wilcoxon signed-rank tests, Holm-corrected (Holm, 1979) across the nine comparisons, and by 95% percentile bootstrap intervals for the mean difference (20,000 resamples). Bike-Sharing uses the same procedure within each budget for the nine comparisons with the methods of Table 2 (100,000 resamples). For ACSIncome, ATF-kNN is compared with each of the other methods by two-sided paired t-tests with t-based 95% intervals, Holm-corrected across all comparisons of the run, within each budget and for the per-repetition difference averaged over budgets. For the synthetic suite, we report the mean world-level paired difference, its 95% percentile bootstrap interval (20,000 resamples of the 30 worlds), and the number of worlds won; the markers in Table 8 use two-sided paired t-tests over the 30 worlds, Holm-corrected across the baselines within each mechanism and budget.

C Additional Results of the Natural-Shift and Synthetic Experiments

C.1 Electricity: Temporal Shift

Table 6 lists every method of the Electricity experiment with standard errors, paired wins of ATF-kNN over 30 repetitions, and Holm-adjusted p-values.

C.2 ACSIncome: Geographic Shift

We predict whether income exceeds \$50k from ten features of the 2018 ACS data (Ding et al., 2021); for binary Y, $m(x) = \mathbb{P}(Y = 1 \mid X = x)$, the squared risk is the Brier score, and the correction is applied to the teacher probability. A four-expert MoE teacher trained and tuned on California is adapted to Michigan with 10,000 unlabeled examples. ATF-kNN reduces the Brier score from 0.14692 to 0.13857 (Table 7), improves on the teacher at every budget, and raises accuracy by 1.66 percentage points with 8 coefficients. It is significantly more accurate than the readout-adapted PP-SSL and SADA, a same-readout RBF residual model, and full fine-tuning, and it is statistically tied with router and expert-head fine-tuning while adapting in 0.11 s on a CPU instead of 3.48–6.21 s for fine-tuning (Table 10). Two-parameter scalar affine correction is slightly more accurate here (0.13496), which indicates that most of the residual is an affine function of the teacher probability; the gap narrows from 0.0068 at $n_s = 10$ to 0.0020 at $n_s = 40$ (Corollary 4).

Table 6: **Electricity, 2012–2013 → 2014,** $n_s = 40$**.** Log-RMSE, mean ± standard error over 30 repetitions; wins of ATF-KNN out of 30; Holm-adjusted paired p-values against ATF-KNN.

Method	log-RMSE (mean ± SE)	ATF-KNN wins (of 30)	Holm-adjusted Wilcoxon p	Trainable parameters
Teacher	$0.4434 \pm 0.0000^{\dagger}$	29	6.5×10^{-7}	0
Mean correction	$0.4480 \pm 0.0015^{\dagger}$	29	2.2×10^{-8}	1
Scalar affine correction	$0.4118 \pm 0.0031^{\dagger}$	27	9.4×10^{-6}	2
Target-only fit	$0.5514 \pm 0.0139^{\dagger}$	30	1.7×10^{-8}	17
PP-SSL (readout)	$0.5926 \pm 0.0083^{\dagger}$	30	1.7×10^{-8}	8
SADA (readout)	$0.8856 \pm 0.0084^{\dagger}$	30	1.7×10^{-8}	8
Router fine-tuning	$0.4154 \pm 0.0019^{\dagger}$	28	2.8×10^{-5}	1,732
Expert-head fine-tuning	0.4006 ± 0.0038	16	0.543	132
ATF-NO-KNN	0.3954 ± 0.0039	8	0.065	8
ATF-KNN	0.3993 ± 0.0038	–	–	8

Table 7: **ACSIncome**, averaged over $n_s \in \{10, 20, 30, 40\}$ (50 repetitions each). †: ATF-KNN significantly better (paired t-test, Holm).

Method	Brier	Accuracy (%)	Params
Teacher	$0.14692^{\dagger}$	78.58	0
Scalar affine	0.13496	80.41	2
Target-only	$0.19652^{\dagger}$	70.36	816
PP-SSL	$0.15495^{\dagger}$	78.43	8
SADA	$0.16089^{\dagger}$	77.11	8
RBF	$0.14587^{\dagger}$	79.28	10–40
Router FT	0.14156	79.86	56,644
Expert-head FT	0.13660	80.06	260
Full FT	$0.14316^{\dagger}$	79.83	265,800
ATF-NO-KNN	$0.14475^{\dagger}$	79.49	8
ATF-KNN	0.13857	80.24	8

C.3 Teacher-Independent Synthetic Suite

We further generate 30 synthetic worlds whose target distribution differs from the source by a covariate shift of a latent six-component mixture and by a change of regression function, $m = m_0 + 0.75\,\tilde{g}$ with a random smooth $\tilde{g}$; both are drawn and fixed before a four-expert top-2 MoE is trained on source data and frozen. With $n_s = 40$ labels and smooth target shifts, ATF-KNN reduces RMSE to 0.7427, compared with 0.7906 for scalar affine correction (Table 8; 95% world-bootstrap interval of the difference $[-0.0775, -0.0212]$, better in 19 of 30 worlds). Permuting the readout coordinates other than $\hat{f}$ jointly across examples raises the RMSE of ATF-NO-KNN from 0.7503 to 0.8505 (worse in 26 of 30 worlds), so the gain depends on the example-level routing. In control worlds with $m = 1.15\,m_0 + 0.25$, the residual is approximately affine in $\hat{f}$ when $\hat{f} \approx m_0$, so $G_{\phi|\mathrm{sc}}$ is small (Theorem 1); scalar correction, which fits fewer coefficients, is then the most accurate, and ATF-KNN still reduces the teacher's error. Expert-head fine-tuning with parameter access is more accurate on this suite (0.7117) but needs about 90 times the adaptation time.

Table 9 reports the smooth mechanism at every label budget.

C.4 Adaptation Cost and Routing Diagnostics

Runtime. Table 10 reports adaptation time and trainable parameters. Times exclude teacher training, which is shared by all methods.

Table 8: **Synthetic suite**, RMSE at $n_s = 40$ (30 worlds); permuted: routed features permuted; –: not run. †: as in Table 7.

Method	Smooth shift	Scalar shift (control)
Teacher	0.8180†	0.4564†
Mean correction	0.8282†	0.4018
Scalar affine	0.7906†	0.3939
Target-only	0.7879	0.7014†
PP-SSL	0.7579	0.4466†
SADA	0.7533	0.4546†
Router FT	0.7825	–
Expert-head FT	0.7117	–
ATF-NO-κNN	0.7503	0.4410†
ATF-NO-κNN, permuted	0.8505†	0.4502†
ATF-κNN	0.7427	0.4280

Table 9: **Synthetic suite, smooth mechanism.** RMSE, mean over 30 worlds.

	Smooth mechanism (mean RMSE over 30 worlds)			
Method	n_s=10	n_s=20	n_s=40	n_s=80
Teacher	0.8180	0.8180	0.8180†	0.8180†
Mean correction	0.8695†	0.8460	0.8282†	0.8212†
Scalar affine correction	0.8805†	0.8280	0.7906†	0.7737†
Target-only fit	1.0533†	0.9341†	0.7879	0.6923
PP-SSL (readout)	0.8968†	0.8072	0.7579	0.7209
SADA (readout)	0.9266†	0.8128	0.7533	0.7163
Router fine-tuning	0.8007	0.7907	0.7825	0.7737†
Expert-head fine-tuning	0.7807	0.7553	0.7117	0.6798
ATF-NO-κNN	0.9167†	0.8013	0.7503	0.7211
ATF-κNN	0.8099	0.7788	0.7427	0.7126

Routing diagnostics. On Electricity, the learned routing is compared with 200 random shuffles of the routing weights on held-out source and target data, and none of the shuffles attains a lower error; all four experts remain active on the deployment year. A permutation test of mean residuals across routing regions gives $p < 0.001$, and a cross-fitted affine correction in the routed readout reaches RMSE 0.3924, compared with 0.4226 for a correction that uses the scalar prediction alone. In the synthetic suite, permuting the routed features while preserving the scalar prediction raises the RMSE of the routed fit from 0.7503 to 0.8505 at $n_s = 40$.

Table 10: **Adaptation cost.** Wall-clock adaptation time and trainable parameters. ACSIncome: all methods on the same CPU. Synthetic suite: ATF-KNN on the CPU and fine-tuning on a GPU of the same machine.

Method	Adaptation time (s)	Trainable parameters
ACSIncome CA→MI: mean over 50 seeds × 4 budgets (CPU)		
Frozen-teacher forward (shared)	0.03	0
ATF-KNN	0.11	8
ATF-KNN + shared teacher forward	0.14	8
Router fine-tuning	3.48	56,644
Expert-head fine-tuning	5.30	260
Full fine-tuning	6.21	265,800
Synthetic, smooth, n_s=40: median over 30 worlds		
ATF-KNN	0.0246	8
Expert-head fine-tuning	2.2470	260

D Additional Empirical Comparisons

To isolate the contribution of each coordinate family in ϕ, Table 11 repeats the synthetic corrupted-oracle sweep while replacing the full feature map by gate-only features $\pi_{1:K-1}(x)$, expert-only features $\mu(x)$, or the scalar teacher feature $\hat{f}(x)$ in the same kNN residual-imputation template. The scalar teacher feature is competitive at the smallest labeled size, reflecting its lower variance. As n_s grows, the full affine teacher-feature map becomes best: gate-only features miss expert-specific gain errors, expert-only features ignore routing, and the no-kNN full-span ablation closes part of the gap but remains behind the full method.

Table 11: **Feature-map ablation on the synthetic corrupted-oracle benchmark.** Test RMSE, mean ± std over 12 seeds. Bold marks the best mean in each column.

Feature/correction variant	$n_s = 20$	$n_s = 40$	$n_s = 80$
Scalar mean correction	0.613 ± 0.039	0.613 ± 0.045	0.598 ± 0.016
Scalar teacher kNN	$\mathbf{0.397 \pm 0.074}$	0.398 ± 0.065	0.352 ± 0.023
Gate-only kNN	0.491 ± 0.066	0.446 ± 0.061	0.395 ± 0.022
Expert-only kNN	0.509 ± 0.044	0.502 ± 0.044	0.446 ± 0.022
Full ϕ, no kNN	0.511 ± 0.069	0.433 ± 0.044	0.361 ± 0.040
Full ϕ + kNN (ATF-KNN)	0.424 ± 0.079	$\mathbf{0.397 \pm 0.044}$	$\mathbf{0.345 \pm 0.034}$

D.1 Additional Baseline and No-kNN Comparisons

Tables 12, 13, and 14 report the full set of standard baselines in the synthetic IID, Diamonds IID, and synthetic clean-shift settings, including MoE-adapted PP-SSL and SADA comparators, together with ATF-NO-KNN, the same affine teacher-feature ridge used by ATF-KNN but trained only on labeled residuals. Unlabeled covariates still set the feature scaling for ATF-NO-KNN, but they do not contribute imputed residual targets. The comparison separates the affine feature contribution from the kNN residual-imputation contribution while keeping the baseline set aligned with the main experiments.

Table 12: **Additional comparison on the synthetic IID corrupted-oracle benchmark.** Test RMSE, mean ± std over seeds. Bold marks the best mean in each labeled-size column.

Method	$n_s = 20$	$n_s = 40$	$n_s = 80$
TEACHER	0.599 ± 0.004	0.599 ± 0.004	0.598 ± 0.004
LABELED-ONLY	1.619 ± 0.169	2.014 ± 0.341	1.391 ± 0.188
Scalar mean correction	0.613 ± 0.039	0.613 ± 0.045	0.598 ± 0.016
Raw-covariate residual	1.013 ± 0.193	1.720 ± 0.357	1.199 ± 0.152
SELF-TRAINING	0.732 ± 0.006	0.729 ± 0.006	0.723 ± 0.007
Scalar teacher mix	0.915 ± 0.174	1.513 ± 0.372	1.112 ± 0.130
MoE-adapted PP-SSL	0.500 ± 0.096	0.462 ± 0.040	0.377 ± 0.045
MoE-adapted SADA	0.598 ± 0.101	0.636 ± 0.101	0.572 ± 0.073
ATF-NO-KNN	0.511 ± 0.069	0.433 ± 0.044	0.361 ± 0.040
ATF-KNN	$\mathbf{0.424 \pm 0.079}$	$\mathbf{0.397 \pm 0.044}$	$\mathbf{0.345 \pm 0.034}$

Table 13: **Additional comparison on Diamonds IID.** Test RMSE, mean ± std over seeds. Bold marks the best mean in each labeled-size column.

Method	$n_s = 20$	$n_s = 40$	$n_s = 80$
TEACHER	0.253 ± 0.054	0.253 ± 0.054	0.253 ± 0.054
LABELED-ONLY	0.908 ± 0.062	0.758 ± 0.043	0.533 ± 0.033
Scalar mean correction	0.248 ± 0.049	0.244 ± 0.046	0.245 ± 0.048
Raw-covariate residual	0.245 ± 0.030	0.254 ± 0.030	0.297 ± 0.049
SELF-TRAINING	0.235 ± 0.073	0.236 ± 0.074	0.233 ± 0.073
Scalar teacher mix	0.260 ± 0.026	0.255 ± 0.030	0.275 ± 0.031
MoE-adapted PP-SSL	0.300 ± 0.030	0.270 ± 0.017	0.251 ± 0.006
MoE-adapted SADA	0.335 ± 0.027	0.320 ± 0.012	0.313 ± 0.005
ATF-NO-KNN	0.235 ± 0.033	0.214 ± 0.027	0.188 ± 0.015
ATF-KNN	$\mathbf{0.227 \pm 0.032}$	$\mathbf{0.213 \pm 0.029}$	$\mathbf{0.187 \pm 0.012}$

Table 14: **Additional comparison under synthetic clean covariate shift.** Test RMSE, mean ± std over seeds. Bold marks the best mean in each labeled-size column. A dagger on the shaded ATF-KNN row indicates that its paired gap from the best mean is not significant at the 5% level across seeds.

Method	$n_s = 20$	$n_s = 40$	$n_s = 80$
TEACHER	1.037 ± 0.298	1.089 ± 0.282	1.103 ± 0.220
LABELED-ONLY	1.781 ± 0.304	1.373 ± 0.335	0.992 ± 0.139
Scalar mean correction	$\mathbf{0.875 \pm 0.212}$	0.857 ± 0.137	0.889 ± 0.120
Raw-covariate residual	0.997 ± 0.281	1.071 ± 0.255	0.971 ± 0.123
SELF-TRAINING	1.042 ± 0.282	1.083 ± 0.272	1.069 ± 0.230
Scalar teacher mix	1.045 ± 0.296	1.094 ± 0.239	0.973 ± 0.127
MoE-adapted PP-SSL	1.163 ± 0.218	1.056 ± 0.190	1.043 ± 0.281
MoE-adapted SADA	1.446 ± 0.493	1.416 ± 0.441	1.474 ± 0.460
ATF-NO-KNN	1.282 ± 0.508	1.461 ± 0.845	0.920 ± 0.159
ATF-KNN	$0.901 \pm 0.221^{\dagger}$	$\mathbf{0.827 \pm 0.147}$	$\mathbf{0.840 \pm 0.169}$

Table 15 adds standard deviations to the expert-number misspecification comparison of Table 4. The second-best method changes across K_{teacher}, but ATF-KNN remains lowest throughout; the table supports robustness to moderate expert-number misspecification rather than a claim that larger K_{teacher} is intrinsically better.

Table 15: **Full expert-number misspecification comparison.** Synthetic shift with $K_{\text{true}} = 3$ and $n_s = 40$. Entries are held-out RMSE mean $\pm$ std over 20 seeds.

	K_{teacher}			
Method	2	3	4	5
Teacher only	1.222 ± 0.189	1.129 ± 0.249	0.901 ± 0.271	0.909 ± 0.286
Self-training	1.206 ± 0.182	1.118 ± 0.238	0.912 ± 0.253	0.924 ± 0.266
Scalar teacher mix	1.218 ± 0.182	1.252 ± 0.203	1.180 ± 0.180	1.161 ± 0.205
MoE-adapted PP-SSL	1.081 ± 0.189	1.086 ± 0.220	0.885 ± 0.195	1.047 ± 0.266
MoE-adapted SADA	1.551 ± 0.472	1.448 ± 0.462	1.185 ± 0.316	1.472 ± 0.463
ATF-KNN	$\mathbf{0.912 \pm 0.107}$	$\mathbf{0.828 \pm 0.117}$	$\mathbf{0.705 \pm 0.143}$	$\mathbf{0.695 \pm 0.147}$

Taken together, Tables 12, 13, 14, and 15 show broad gains. At $n_s = 20$ under synthetic shift, the scalar mean correction has the best mean and ATF-KNN is statistically tied by the paired test. In the other labeled-size comparisons and in every wrong-K column, ATF-KNN has the best mean. ATF-NO-KNN is close to ATF-KNN on Diamonds and at larger n_s, but it is worse in the scarce-label synthetic settings, supporting the interpretation that unlabeled residual imputation is most useful when labeled residuals alone under-identify the affine teacher-feature correction.

E Experimental Settings and Sensitivity

E.1 Benchmark configurations

Synthetic corrupted-oracle. Synthetic MoE generator with $K_{\text{true}} = 3$ soft-partitioning experts over d=8 features. The teacher is pretrained on n_{pre}=2000 samples and then frozen. Post-pretraining corruptions: gate mixing ρ=0.5, expert bias scale b=0.5. We sweep $n_s \in \{20, 40, 80\}$ with n_u=5000, n_e=4000, 12 seeds.

Diamonds. Diamonds, 54,000 rows. The teacher is a linear MoE with K=5 experts defined by regimes over (`carat`,`cut`), pretrained on n_{pre}=12,000. Corruptions: pretrain-keep $(1.0, 0.9, 0.45, 0.15, 0.05)$, gate mixing ρ=0.05, expert bias $(0.0, 0.04, 0.1, 0.2, 0.34)$, expert noise $(0.0, 0.0, 0.02, 0.04, 0.07)$. Evaluation uses n_e=6000; we sweep $n_s \in \{20, 40, 80\}$ with n_u=6000, 6 seeds.

Synthetic pretrain corruption. Same generator as the synthetic corrupted-oracle but with corruption moved into pretraining: pretrain-keep $(1.0, 0.5, 0.15)$, mask scale $(0.0, 0.40, 0.70)$, noise scale $(0.0, 0.05, 0.10)$. n_s=40, n_u=5000, 12 seeds.

Synthetic non-affine teacher stress. Same generator as the synthetic corrupted-oracle, but the exposed oracle teacher components are distorted by per-expert nonlinear terms: $\tanh(0.9\mu_1)^2 - 0.45$, $\sin(1.1\mu_2)$, and $\tanh(\mu_3)(0.5 + 0.25x_6^2)$, scaled by 0.45, with gate mixing ρ=0.20. This makes the teacher residual only partially captured by the affine ϕ span. n_s=40, n_u=5000, n_e=4000, 12 seeds.

Bike-Sharing with neural MoE. Small neural MoE (2-layer MLP experts and gate, hidden width 32) on UCI Bike-Sharing, with K=5 regimes induced by (`hour`,`weather`). Pretraining uses n_{pre}=9000 with pretrain-keep $(1.0, 1.0, 0.25, 0.12, 0.05)$ and feature corruption, so the teacher is imperfect on its own distribution. The label budgets, repetitions, and fine-tuning controls of the experiment in Section 5.1 are given in Appendix B.

Synthetic clean covariate shift. Same generator as the synthetic corrupted-oracle, with clean covariate shift added: $p(Y \mid X)$ preserved, $p(X)$ perturbed by mean-shift δ=1.5 toward rarer regimes. Pretrain-keep $(1.0, 0.5, 0.15)$, expert dim 6, gate mixing ρ=0.2. $n_s \in \{20, 40, 80\}$, n_u=5000, n_e=4000, 8 seeds.

Diamonds deployment-shift stress. Real-data deployment stress on Diamonds with source/deployment mismatch: target bias scales $(0.0, 0.0, 0.08, -0.22, 0.38)$, pretrain-keep $(1.0, 0.7, 0.15)$, ρ=0.06, expert bias $(0.0, 0.03, 0.1)$. $n_s \in \{20, 40\}$, n_u=6000, 12 seeds.

Wrong-K sweep. We fix K_{true}=3 and sweep $K_{\text{teacher}} \in \{2, 3, 4, 5\}$, retraining the teacher for each value. Other settings follow the clean-shift configuration (δ=1.5, n_s=40, n_u=5000), 20 seeds per cell. A no-mixing support setting additionally runs K_{teacher}=5 vs. K_{true}=3 with 12 seeds.

E.2 Fashion-MNIST Classification Support

We include a binary image-classification support experiment to check whether the same gray-box MoE readout is useful beyond squared-error regression. This experiment is not used as a main benchmark because the method and theory are written for conditional-mean estimation, but it tests the same operational question: whether routing probabilities and expert-specific predictions carry adaptation signal that is lost by treating the teacher as a scalar probability source.

Problem setup. The task is Fashion-MNIST upper-vs-rest classification: $Y = 1$ for upper-body clothing labels $\{0, 2, 3, 4, 6\}$, corresponding to T-shirt/top, pullover, dress, coat, and shirt, and $Y = 0$ for the remaining categories. Raw 28×28 images are projected to 32 PCA features fit on the pretraining, labeled, and unlabeled covariates for each seed. Each run uses $n_{\text{pre}} = 500$ teacher-pretraining examples, $n_s = 40$ labeled deployment examples, $n_u = 3000$ unlabeled deployment covariates, and $n_e = 2000$ held-out evaluation examples, averaged over 12 seeds.

Expert setup. The teacher is a federated-domain MoE classifier with $K = 5$ experts, where each expert is trained on one synthetic pretraining domain. The domains are formed class by class. For each original Fashion-MNIST class $c \in \{0, \ldots, 9\}$, we sample domain proportions

$$(p_{c1}, \ldots, p_{c5}) \sim \text{Dirichlet}(0.4, 0.4, 0.4, 0.4, 0.4),$$

and then split the pretraining images from class c across the five domains according to these proportions. When enough examples are available, we first place at least five images of class c in each domain so that every expert sees every original class at least minimally. The parameter $0.4 < 1$ makes the sampled proportions uneven: for a given class, most examples are usually assigned to one or two domains, while other domains receive only a small share. After repeating this procedure for all ten classes, each domain contains examples from all classes but with a different class composition; this creates heterogeneous, federated-style expert training data rather than five i.i.d. copies of the same pretraining set.

The MoE teacher is then fit from these domain-labeled pretraining data. A multinomial logistic gate is trained to predict the domain index from PCA features, and expert k is a binary logistic classifier trained only on the examples assigned to domain k. To make the teacher informative but imperfect, the binary pretraining labels used to fit the experts are flipped independently within each domain at either 5% or 15% before fitting; the gate/expert model is then held fixed. At adaptation time the gray-box readout is $\pi(x)$ and expert probabilities $q_k(x)$, with teacher probability $\hat{p}(x) = \sum_k \pi_k(x) q_k(x)$. ATF-kNN uses the same affine readout coordinates $[\pi_{1:K-1}, \pi_1 q_1, \ldots, \pi_K q_K]$ and fits a residual update for probabilities.

Evaluation metrics. We report proper scoring rules rather than threshold accuracy. For predicted probability $\tilde{p}_i$, the Brier score is $n_e^{-1} \sum_i (y_i - \tilde{p}_i)^2$, and cross entropy is

$$-\frac{1}{n_e} \sum_{i=1}^{n_e} \{y_i \log \tilde{p}_i + (1 - y_i) \log(1 - \tilde{p}_i)\}.$$

Lower values are better for both metrics. The table reports mean $\pm$ standard deviation over seeds; the full per-seed rows and additional baselines are included in the supplementary material.

ATF-kNN improves Brier score by 12.7% and cross entropy by 14.5% relative to the teacher under 5% teacher-label noise. Under 15% teacher-label noise, the improvements increase to 18.8% and 19.8%, respectively. PP-SSL and SADA probability-score baselines are less effective than ATF-kNN under both proper scoring rules, suggesting that probability aggregation alone does not fully capture

Table 16: Fashion-MNIST upper-vs-rest support experiment. The result supports probability-estimation improvement under an imperfect gray-box MoE classifier; it is not used to claim threshold-classification dominance.

Teacher label noise	Method	Brier ↓	Cross entropy ↓
5%	Teacher	0.0556 ± 0.0054	0.2173 ± 0.0156
5%	Mean correction	0.0557 ± 0.0054	0.2178 ± 0.0155
5%	Self-training	0.0655 ± 0.0040	0.2507 ± 0.0128
5%	PP-SSL	0.2081 ± 0.0284	0.8677 ± 0.1532
5%	SADA-style aggregation	0.1248 ± 0.0379	0.4985 ± 0.2553
5%	ATF-KNN	$\mathbf{0.0485 \pm 0.0058}$	$\mathbf{0.1857 \pm 0.0177}$
15%	Teacher	0.0871 ± 0.0052	0.3155 ± 0.0149
15%	Mean correction	0.0874 ± 0.0051	0.3156 ± 0.0149
15%	Self-training	0.0889 ± 0.0038	0.3238 ± 0.0129
15%	PP-SSL	0.2106 ± 0.0215	0.9225 ± 0.1191
15%	SADA-style aggregation	0.1213 ± 0.0306	0.4155 ± 0.0713
15%	ATF-KNN	$\mathbf{0.0708 \pm 0.0060}$	$\mathbf{0.2529 \pm 0.0190}$

the routed residual structure. This pattern is consistent with the regression experiments: when the deployed teacher is informative but its errors vary across routed states, the expert readout provides a useful adaptation coordinate system.

E.3 Hyperparameters and CV protocol

Synthetic IID. The synthetic corrupted-oracle and synthetic pretrain-corruption runs use fixed method defaults: polynomial ridge baselines use α=1.0, while scalar teacher mix and ATF-KNN use α=1.5. ATF-KNN uses k_{nn}=5 neighbors and unlabeled weight w_u=0.5.

Real IID and shift experiments. Diamonds IID, Bike-Sharing IID, synthetic shift, wrong-K, and real-shift runs tune α by 3-fold CV on the labeled set for every ridge-based method using the grid $\{0.25, 0.5, 1.0, 1.5, 2.0, 4.0\}$. Unless otherwise specified, ATF-KNN uses k_{nn}=5 and w_u=0.5. The Bike-Sharing neural-MoE support run uses k_{nn}=15 for ATF-KNN. The additional comparisons in Appendix D.1 use the same seeds and tuning rules as the corresponding benchmark cells, with ATF-NO-KNN added to isolate the kNN residual-imputation step.

E.4 Hyperparameter Sensitivity

We conduct ablation studies on the two non-ridge ATF-KNN hyperparameters in the synthetic corrupted-oracle benchmark at n_s=40, using the same 12 seeds as the main IID sweep. The ridge penalty α is not included in this sweep because it is handled by the fixed protocol or labeled-set CV described above.

Table 17 varies k_{nn}, the number of nearest neighbors in the kNN residual imputer, while fixing the effective unlabeled weight at the default value. Moderate neighbor counts, i.e., moderate values of k_{nn} in the kNN imputer, are stable: k_{nn}=5 and 10 are essentially tied, and $k_{\mathrm{nn}} \in \{3, 15, 20\}$ still improves over the no-kNN ablation. The one-neighbor imputer is too noisy and falls behind the no-kNN fit.

Table 17: **Sensitivity to the kNN neighbor count.** Synthetic corrupted-oracle benchmark at n_s=40, 12 seeds. Values are RMSE mean $\pm$ SE.

Variant	k_{nn}	RMSE
ATF-KNN	1	0.4420 ± 0.0286
ATF-KNN	3	0.4076 ± 0.0164
ATF-KNN	5	$\mathbf{0.3972 \pm 0.0126}$
ATF-KNN	10	0.3976 ± 0.0129
ATF-KNN	15	0.4034 ± 0.0158
ATF-KNN	20	0.4116 ± 0.0165
ATF-NO-KNN	–	0.4329 ± 0.0127
Teacher only	–	0.5986 ± 0.0011

Table 18 varies the aggregate unlabeled weight through $\rho_u = w_u n_u / n_s$, which is the effective unlabeled-to-labeled weight ratio in the ridge objective. This scale is more interpretable than the raw per-sample weight w_u because n_u is much larger than n_s. Performance improves smoothly as ρ_u moves away from zero and saturates around $\rho_u = 5$ to 10; the reported default w_u=0.5 corresponds to $\rho_u = 62.5$ in this setting and lies on the same plateau. Thus the important distinction is between using enough imputed residual mass and using none, not fine-tuning the exact positive weight.

Table 18: **Sensitivity to the effective unlabeled weight.** Synthetic corrupted-oracle benchmark at n_s=40, n_u=5000, 12 seeds, with k_{nn}=5. The effective ratio is $\rho_u = w_u n_u / n_s$. Values are RMSE mean $\pm$ SE.

ρ_u	Equivalent w_u	RMSE
0	0.000	0.4329 ± 0.0127
0.25	0.002	0.4185 ± 0.0118
0.5	0.004	0.4117 ± 0.0116
1	0.008	0.4051 ± 0.0118
2	0.016	0.4005 ± 0.0121
5	0.040	0.3978 ± 0.0124
10	0.080	0.3972 ± 0.0125
62.5	0.500	$\mathbf{0.3972 \pm 0.0126}$

F BASELINE IMPLEMENTATION DETAILS

Table 19: **Baselines.** "Uses $\mathcal{U}$ signal" means unlabeled points enter as pseudo-labels, prediction-powered gradients, aggregation moments, or imputed residuals, not merely as a feature-scaling reference. "MoE readout" records whether the method consumes expert outputs or both routing probabilities and expert outputs beyond the scalar $\hat{f}$.

Method	Uses $\mathcal{U}$ signal	MoE readout	Baseline idea
Teacher only	–	$\hat{f}$ only	use the frozen teacher output directly
Labeled only	–	$\hat{f}$ only	supervised polynomial ridge on labeled data
Scalar mean correction	–	$\hat{f}$ only	add the labeled mean residual to the teacher
Scalar affine correction	–	$\hat{f}$ only	ridge fit of labeled residuals on $[1, \hat{f}]$
Raw-covariate residual	–	$\hat{f}$ only	fit teacher residuals in raw covariate space
Self-training	✓	$\hat{f}$ only	refit on labeled data plus teacher pseudo-labels
Scalar teacher mix	–	$\hat{f}$ only	combine scalar teacher prediction with a labeled residual fit
MoE-adapted PP-SSL (Shoham et al., 2025)	✓	(π, μ)	prediction-powered gradient on an affine MoE-feature student
MoE-adapted SADA (Shan et al., 2026)	✓	(π, μ)	estimating-equation aggregation with affine MoE features and scalar/expert predictions
RBF (readout)	–	(π, μ)	RBF-kernel residual model on the affine MoE features
Router / expert-head / full fine-tuning	–	parameters	L2-SP update of the router, the experts' final layers, or all parameters
ATF-KNN (ours)	✓	(π, μ)	residual ridge in affine MoE-feature space with kNN-imputed unlabeled residuals

Scalar mean correction. Implements the zero-bias scalar correction $\hat{y}(x) = \hat{f}(x) + \bar{r}_s$ where $\bar{r}_s$ is the mean of labeled residuals. It is a simple point-prediction analogue of residual mean adjustment ideas in semi-supervised correction (Zhang et al., 2019), but it uses no unlabeled training signal.

Scalar affine correction. Fits $\hat{f} + a + b\hat{f}$ by ridge regression on the labeled residuals; it spans $\mathcal{H}_{\text{sc}}$ and is therefore nested in the ATF-KNN correction class.

Raw-covariate residual. Implements a degree-2 polynomial ridge fit on labeled residuals (x_s, r_s), with $\mathcal{S} \cup \mathcal{U}$ used only to define the feature standardization scale. It is a raw-covariate residual-regression reference point, not a full unlabeled-reweighted implementation of the semi-supervised linear-regression estimator of Chakrabortty & Cai (2018).

Self-training. Degree-2 polynomial ridge fit on the concatenated dataset $(\mathcal{S} \cup \mathcal{U},\ y_s \cup \hat{f}(x_u))$, with unlabeled targets taken from the teacher.

Scalar teacher mix. Scalar teacher/labeled-fit mixture: the mixing weight is chosen by labeled-fold CV, and the residual component is fit in raw covariate space. No gate or per-expert readout is used beyond the scalar $\hat{f}$.

MoE-adapted PP-SSL. We adapt the online prediction-powered SSL update of Shoham et al. (2025) to the gray-box MoE setting by using the affine MoE readout $[1, \phi(x)]$ as the student covariate vector. The regression implementation applies the prediction-powered gradient to this linear readout student, with the scalar teacher prediction $\hat{f}$ supplying pseudo-labels in the unlabeled gradient term. This baseline tests whether the MoE readout coordinates alone are sufficient when inserted into a generic prediction-powered objective; it does not use kNN-imputed residuals or the proposed residual-ridge objective.

MoE-adapted SADA. We adapt the estimating-equation aggregation idea of Shan et al. (2026) to the gray-box MoE predictor by using the affine MoE readout $[1, \phi(x)]$ as the regression design and

$[\hat{f}, \mu_1, \ldots, \mu_K]$ as prediction sources in the aggregation moments. The reported regression baseline uses fixed regularization and shrinkage constants chosen before the final full reruns. This is an adapted comparator for multi-prediction estimating equations in the exposed MoE feature space, not a claim that the original SADA method was designed for MoE readouts.

RBF (readout). An RBF-kernel ridge model of the labeled residuals on the same standardized affine MoE features used by ATF-KNN; it tests whether a nonlinear residual model in the readout helps at these label budgets.

Fine-tuning controls. Router, expert-head, and full fine-tuning update the corresponding teacher parameters on the labeled deployment data with an L2-SP penalty (Li et al., 2018) toward the pretrained values, with hyperparameters selected by cross-validation on the labeled set (Appendix B). They require access to the weights and the training pipeline and serve as stronger-access references.

ATF-NO-KNN. Ridge on $(\tilde{\phi}_s, r_s)$ only: the ATF-KNN affine teacher-feature correction with the kNN residual-imputation step removed. The unlabeled covariates still define the standardization scale for ϕ, but they do not add imputed residual targets. This ablation isolates the contribution of unlabeled residual imputation beyond the affine teacher-feature span.

Labeled only. Degree-2 polynomial ridge on (x_s, y_s), ignoring $\mathcal{U}$ as a training signal and ignoring $\hat{f}$; a pure small-sample supervised baseline. In the natural-shift, synthetic-suite, and Bike-Sharing experiments the corresponding baseline is called target-only fit; its model class in each experiment is given in Appendix B.

G Reproduction Protocol

Code and raw outputs are provided in the supplementary material. Each benchmark uses a deterministic base seed offset; figures and tables state whether they report standard error or standard deviation, and the raw per-seed rows are included in the supplementary material, together with scripts that regenerate the tables from them. Dependencies: Python 3.9 (3.10 for the synthetic suite) with standard scientific stack (`numpy`, `pandas`, `scipy`, `scikit-learn`, `torch`, `torchvision`) and `folktables`. Datasets used are the UCI Electricity Load Diagrams (Trindade, 2015), the 2018 ACS PUMS files accessed through `folktables` (Ding et al., 2021), Diamonds, UCI Bike-Sharing (Fanaee-T & Gama, 2014), Fashion-MNIST (Xiao et al., 2017) via `torchvision`, and synthetic data generated according to the specifications in Appendices B and E.1. The natural-shift, Bike-Sharing, and controlled experiments run on a CPU; the synthetic suite, which trains 30 teachers, was run on a single consumer GPU.