ADAPTIVE TRANSFER OF REPRESENTATIONAL SIGNATURES FOR LLM RELIABILITY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are reported to be susceptible to reliability issues including hallucinations and reasoning errors. Existing quality detectors can predict unreliable outputs based on internal representations with decent accuracies. However, training these detectors often requires substantial domain-specific labeled data, which limits their applicability to new domains. We propose to address this challenge by transferring representational signatures across related domains. More specifically, we build a statistical framework that first extracts low-dimensional signatures for each domain, then uses cross-fitting to learn composition weights of domain-specific predictors. The weights adapt to each source domain's predictive relevance to the target domain. Our asymptotic analysis of a least-squares analogue characterizes the limiting prediction risk and aggregation weights in terms of domain discrepancies and representation bias. Experiments across multiple hallucination datasets and open-weight LLMs show that, with only 25 labeled questions in the new domain, adaptive transfer improves hallucination-detection AUROC by 7.8%. For reasoning-step prediction, it improves AUROC by 7.7%.

1 INTRODUCTION

Large language models (LLMs) have demonstrated strong performance across a wide range of tasks, including language generation (Brown et al., 2020), mathematical reasoning (Lewkowycz et al., 2022), and code generation (Chen et al., 2021). However, strong performance on these tasks does not necessarily imply reliable behavior. Hallucinations and misalignment can lead to incorrect decisions (Hager et al., 2024; Hu et al., 2026b) or unintended actions (Zhang et al., 2025a). In July 2026, for example, OpenAI reported that models undergoing internal cybersecurity evaluations with reduced safeguards bypassed containment controls and compromised Hugging Face's production infrastructure (OpenAI, 2026). These risks highlight the need to detect unreliable LLM behaviors.

Recent studies have linked unreliable behavior in open-weight LLMs to specific patterns in their internal representations. For hallucination detection, for example, linear probes trained on hidden-state activations from a single layer can distinguish correct answers from hallucinated ones with high accuracy (O'Neill et al., 2025; Orgad et al., 2025). Other approaches reduce the dimensionality of internal representations before classification (Hu et al., 2026a), or capture more complex patterns using nonlinear classifiers, such as multilayer perceptrons (MLPs) (Azaria & Mitchell, 2023; Zhou et al., 2025) and vision transformers (ViTs) (Bar-Shalom et al., 2025). Together, these methods suggest that internal representations provide informative signals for hallucination detection.

Despite their success, many of these methods rely on labeled examples from a specific dataset or domain (Shamsi, 2026). In practice, labeled examples can be costly to obtain in sufficient quantities for every new domain. Existing work addresses this by reshaping internal states to make truthfulness signals more consistent across domains (Zhang et al., 2025b) or by constructing hallucination scores from uncertainty fluctuations that remain stable under domain shift (Deng et al., 2026). However, transfer performance can vary substantially across domains. A detector trained on a single dataset may generalize better to some related domains while still performing poorly on others (Alansari et al., 2026). This raises a natural question:

Can we adaptively recompose reliability detectors trained on multiple related source domains to achieve accurate detection in a target domain with limited labeled data?

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.

To answer this question, we develop an adaptive transfer method that learns a target-domain detector using information from multiple source domains. For each source domain, we train a classifier on a low-dimensional principal component subspace that captures the dominant variation in the model's internal representations from perturbed prompts. We then linearly combine these source detectors with a detector trained on the available labeled target-domain data. Because the source detectors may differ in their predictive relevance to the target domain, we use cross-fitting to learn composition weights that balance information transferred from the source domains with the information learned directly from the target domain. We call the resulting method Adaptive Signature Transfer (AdaST).

To understand when source-domain information benefits AdaST, we develop an asymptotic analysis of the adaptive transfer procedure in the regression setting. The analysis characterizes how the learned composition weights converge to the population-optimal counterparts, which are dependent on the discrepancy between source and target domains. The theoretical analysis confirms that AdaST asymptotically performs better than or as well as the predictor that only relies on the target domain information.

We use two tasks to validate the numerical performance of AdaST, including hallucination detection and mathematics reasoning correctness prediction. AdaST achieves robust predictive performance across multiple domains on multiple open-weight models. Furthermore, we show that the learned detector vectors can be used as steering directions. Our main contributions are summarized as:

1. We develop an adaptive cross-domain transfer method, AdaST, for representational signatures of LLM behavior. The method constructs principal-component detectors in each domain and uses cross-fitting to learn adaptive transfer weights.
2. We characterize the limiting behavior of the adaptive transfer mechanism in the regression setting, which validates the effectiveness of adaptive transfer when the target domain is similar to the source domains.
3. We demonstrate empirically that internal-state detectors can benefit from cross-domain transfer, with larger gains when fewer labeled examples are available in the target domain.

1.1 Related Work

Representation-based Reliability Detection. Existing representation-based LLM reliability detection methods have explored a variety of designs, including linear probes applied directly to model activations (O'Neill et al., 2025; Orgad et al., 2025) and low-dimensional subspace methods that extract compact behavioral representations (Hu et al., 2026a). More flexible neural classifiers have also been investigated, such as MLPs trained on hidden state activations (Azaria & Mitchell, 2023; Zhou et al., 2025) and Vision Transformers that model the structured layer-by-token organization of activation tensors (Bar-Shalom et al., 2025). These methods can achieve strong performance when sufficient labeled data are available, but their effectiveness can degrade under label scarcity (Shamsi, 2026).

Cross-Domain Reliability Detection. Recent work examines whether reliability signals learned from internal representations can directly transfer across domains. Orgad et al. (2025) report substantial cross-dataset degradation, while Alansari et al. (2026) find that transfer can succeed but varies with the training and test datasets. Other work improves cross-domain robustness by modifying internal representations or constructing more stable detection signals (Zhang et al., 2025b; Hu et al., 2026a; Deng et al., 2026). These methods train a detector on a single source domain and apply it unchanged to a new domain. Our method instead studies how detectors learned from multiple source domains can be adaptively combined for a new domain with limited labeled data.

Multi-Source Transfer Learning. Multi-source transfer learning studies how knowledge from several related source domains can be combined to improve prediction in a target domain. Early studies formalized the multi-source adaptation problem and developed principled ways to combine source-specific predictors (Mansour et al., 2008; 2009). Subsequent methods have explored a variety of strategies for source combination, including mixture-of-experts models with target-dependent weights (Guo et al., 2018), representation alignment (Peng et al., 2019), and stacking or similarity-based weighting (Wu et al., 2023). We build on these multi-source transfer learning ideas to study adaptive transfer across representation-based LLM reliability detectors.

2 METHOD

In this section, we present the proposed AdaST framework for cross-domain binary behavioral prediction, as illustrated in Figure 1. First, for each domain, we query the LLM with multiple semantically equivalent paraphrases of the same questions and extract the final-token hidden states as representations of the generated responses. We then construct question-level contrast vectors by comparing hidden states associated with the two behavioral classes and use these contrasts to estimate a low-dimensional behavioral subspace and train a domain-specific detector. Second, the learned detectors are adaptively combined using weights estimated from cross-fitted predictions in the target domain.

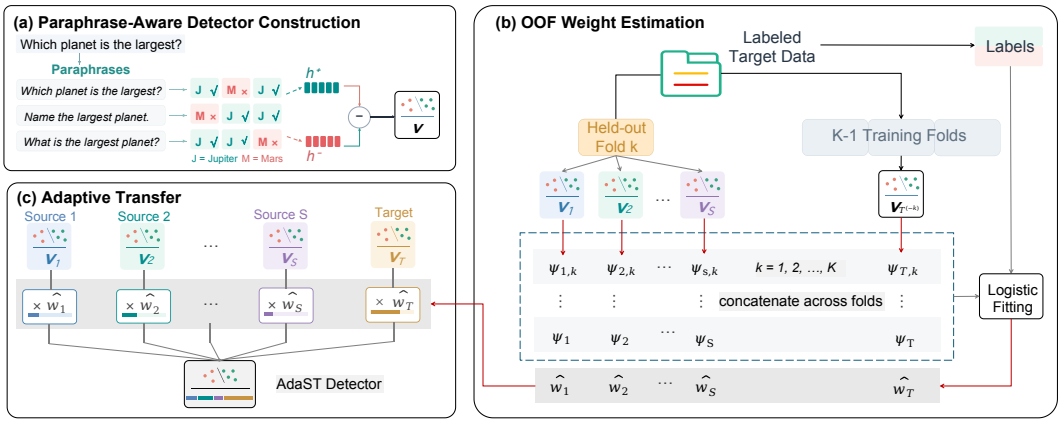


Figure 1: Overview of the proposed AdaST framework.

2.1 PARAPHRASE-BASED BEHAVIORAL SUBSPACE

Let $[S] = \{1, \ldots, S\}$ index the source domains and T index the target domain. For each domain index $s \in [S] \cup \{T\}$, let $\mathcal{Q}_s$ denote the set of questions from the corresponding domain. For example, $\mathcal{Q}_1$ may contain factual question-answering prompts such as "What is the capital of France?", while $\mathcal{Q}_2$ may contain conversational question-answering prompts such as "Which city do you recommend visiting in France?" The semantic similarity across these domains allows us to transfer knowledge about LLM representations.

Fix a domain index $s \in [S] \cup \{T\}$. For each question $q \in \mathcal{Q}_s$, we retain the original wording of q and generate $P - 1$ semantically equivalent paraphrases, inspired by Song et al. (2026). For example, the question "What is the capital of France?" can be paraphrased as "Name the capital of France." We then query the LLM R times for each version of the question q to obtain R responses per version. For each generated response, we extract the hidden state at the final generated token from a selected model layer and denote it by $\mathbf{h}_{qpr} \in \mathbb{R}^D$, where D is the hidden-state dimension, p indexes the question version, and r indexes the generated response. Each response has a binary behavioral label $y_{qpr} \in \{0, 1\}$, where $y_{qpr} = 1$ denotes the positive class and $y_{qpr} = 0$ the negative class. Let $\mathcal{Q}_s^{\pm} \subseteq \mathcal{Q}_s$ denote the questions for which both behavioral classes are observed, and let $m_s = |\mathcal{Q}_s^{\pm}|$.

Following prior work on contrastive representation learning (Zou et al., 2023; Li et al., 2023b; Rimsky et al., 2024), we represent the behavioral difference for each question by comparing the mean representations associated with its positive and negative responses. Let $\mathcal{I}_q^+ = \{(p, r) : y_{qpr} = 1\}$ and $\mathcal{I}_q^- = \{(p, r) : y_{qpr} = 0\}$ denote the positive and negative classes for question q, respectively. For each $q \in \mathcal{Q}_s^{\pm}$, we define the contrast vector:

$$\boldsymbol{\delta}_q = \frac{1}{|\mathcal{I}_q^+|} \sum_{(p,r) \in \mathcal{I}_q^+} \mathbf{h}_{qpr} - \frac{1}{|\mathcal{I}_q^-|} \sum_{(p,r) \in \mathcal{I}_q^-} \mathbf{h}_{qpr}. \tag{1}$$

Here, $\boldsymbol{\delta}_q$ quantifies the representation difference between the positive and negative behavioral classes for the same question q. We stack the contrast vectors as rows of $\mathbf{C}_s$, define their mean as $\bar{\boldsymbol{\delta}}_s = \frac{1}{m_s}\sum_{q\in\mathcal{Q}_s^{\pm}} \boldsymbol{\delta}_q$, and center them across questions to obtain $\mathbf{C}_s - \mathbf{1}\bar{\boldsymbol{\delta}}_s^\top$. Since m_s may be small relative to the hidden-state dimension D, the problem becomes overparameterized. Consequently, fitting a detector directly in the full residual space can be unstable and prone to overfitting. We therefore perform principal component analysis on the centered contrast matrix to obtain a lower-dimensional representation:

$$\boldsymbol{C}_s - \mathbf{1}\bar{\boldsymbol{\delta}}_s^\top = \boldsymbol{L}_s\boldsymbol{\Lambda}_s\boldsymbol{V}_s^\top, \qquad \boldsymbol{B}_s = \boldsymbol{V}_{s,1:k_s}^\top \in \mathbb{R}^{k_s\times D}, \tag{2}$$

We retain $k_s = \min\{k_{\max}, m_s - 1, D\}$ principal directions, where $k_{\max}$ is selected using held-out validation data, and the rows of $\boldsymbol{B}_s$ span the domain-specific behavioral subspace.

For a hidden state $\mathbf{h} \in \mathbb{R}^D$, we first project it onto the behavioral subspace $\boldsymbol{B}_s$. Let $\boldsymbol{\mu}_s$ and $\boldsymbol{\gamma}_s$ denote the coordinate-wise mean and standard deviation of $\boldsymbol{B}_s\mathbf{h}$ over the training responses in domain s. For numerical stability, we standardize the projected representations coordinate-wise and define

$$\boldsymbol{\phi}_s(\mathbf{h}) = \mathrm{diag}(\boldsymbol{\gamma}_s)^{-1}(\boldsymbol{B}_s\mathbf{h} - \boldsymbol{\mu}_s). \tag{3}$$

We then fit a logistic regression in this standardized subspace and denote its coefficient vector by $\boldsymbol{\beta}_s$. The same procedure is applied independently to each source domain and the target domain.

2.2 Adaptive Cross-Domain Transfer

When labeled data in a target domain are scarce, a detector trained only on the target domain may generalize poorly to unseen examples. Meanwhile, labeled datasets from related domains are often available. When the underlying tasks are similar, the corresponding internal representations, here taken as residual-stream activations, may share structure across domains. We therefore supplement the target detector with detectors learned from the source domains.

The domain-specific detectors are learned in different low-dimensional coordinate systems induced by their respective bases $\boldsymbol{B}_s$. Since these coordinate systems are not directly comparable, we map each domain-specific coefficient vector $\boldsymbol{\beta}_s$ back to the shared residual space. Using the standardized projection in Equation 3, the coefficient on $\mathbf{h}$ is

$$\mathbf{v}_s = \boldsymbol{B}_s^\top \mathrm{diag}(\boldsymbol{\gamma}_s)^{-1}\boldsymbol{\beta}_s. \tag{4}$$

Since the detector scores are standardized over the labeled target responses before combination, the offset has no effect on transfer. Only the direction of $\mathbf{v}_s$ matters after this standardization, so we normalize it as $\mathbf{v}_s \leftarrow \mathbf{v}_s/\|\mathbf{v}_s\|_2$ and use $\mathbf{v}_s^\top\mathbf{h}$ to score target hidden states.

The target detector is fitted using the labeled target data, so its predictions on the same examples are in-sample. Using these predictions to learn the combination weights can therefore favor the target detector over the source detectors. To remove such an advantage, we use a K-fold cross-fitting scheme. Specifically, let n_T denote the number of labeled target questions available for training. These questions are partitioned into K equal folds. For each held-out fold, the target subspace and detector are refitted on the other $K-1$ folds, and the corresponding residual-space direction $\mathbf{v}_T$ is recomputed, while the source detectors remain fixed. For each held-out fold k, let $\mathrm{Std}_{-k}(\cdot)$ denote standardization using the other $K-1$ target folds. We apply Std_{-k} to the detector scores $\mathbf{v}_s^\top\mathbf{h}$ on the held-out fold and concatenate the standardized scores over all K folds to form the cross-fitted score vector $\boldsymbol{\psi}_s$. These score vectors are concatenated as

$$\boldsymbol{\Psi} = [\boldsymbol{\psi}_1 \quad \cdots \quad \boldsymbol{\psi}_S \quad \boldsymbol{\psi}_T].$$

Let $\mathbf{w} = (w_1, \ldots, w_S, w_T)^\top$ denote the combination weights, where $w_1, \ldots, w_S$ correspond to the source detectors and w_T to the target detector, and let b denote the intercept. Let i index the labeled target responses (q, p, r). For each response i, let y_i be its binary label and $\boldsymbol{\Psi}_i$ the corresponding row of $\boldsymbol{\Psi}$. We estimate the combination weights and intercept once from the complete $\boldsymbol{\Psi}$ by solving the constrained logistic regression problem:

$$(\widehat{b}, \widehat{\mathbf{w}}) = \arg\min_{\substack{b\in\mathbb{R},\ \mathbf{w}\geq 0 \\ \mathbf{1}^\top\mathbf{w}=1}} \sum_i \left[\log(1+\exp(b+\boldsymbol{\Psi}_i\mathbf{w})) - y_i(b+\boldsymbol{\Psi}_i\mathbf{w})\right]. \tag{5}$$

Equation 5 learns the transfer allocation weights from the cross-fitted scores. The learned weights determine how much each detector contributes to the target prediction, while the intercept adjusts the overall offset of the combined score. For prediction, we train the target detector on all labeled target questions so that all available target supervision is used for prediction. For the fixed source detectors and the trained target detector, we compute the score standardization statistics using all labeled target responses and denote the standardized score for a hidden state $\mathbf{h}$ by $\psi_s(\mathbf{h})$. For a new target response, we combine the standardized scores from the fixed source detectors and the trained target detector using the learned weights. This defines the AdaST detector:

$$\widehat{f}(\mathbf{h}) = \widehat{b} + \sum_{s\in[S]\cup\{T\}} \widehat{w}_s \psi_s(\mathbf{h}). \tag{6}$$

$\widehat{f}(\mathbf{h})$ is a scalar logit for the new target response. Applying the sigmoid function to $\widehat{f}(\mathbf{h})$ gives the predicted probability that the response belongs to the positive class. Larger values therefore indicate stronger evidence for the positive class, whereas smaller values indicate stronger evidence for the negative class.

3 Asymptotic Analysis

3.1 Risk Analysis

To characterize the geometry of transfer, we study a least-squares analogue of the detector in Section 2, using fixed prediction subspaces and convex combinations of unstandardized linear predictions. This formulation characterizes the limiting weights and prediction risk in terms of domain discrepancy and representation bias.

For each domain $s \in [S]\cup\{T\}$ with n_s samples, consider the linear model $y_s = \boldsymbol{x}_s^\top \boldsymbol{\theta}_s^\star + \varepsilon_s$, where $\boldsymbol{x}_s \sim \mathcal{N}(0, \boldsymbol{\Sigma}_s)$ and ε_s is independent noise with mean zero and variance σ_s^2. Let the orthonormal columns of $\boldsymbol{U}_s \in \mathbb{R}^{D\times k}$ span a fixed prediction subspace, which corresponds to $\boldsymbol{B}_s^\top$ in Section 2. The population projected coefficient and the matrix of prediction discrepancies relative to the target are

$$\boldsymbol{\theta}_s^\circ = \boldsymbol{U}_s(\boldsymbol{U}_s^\top \boldsymbol{\Sigma}_s \boldsymbol{U}_s)^{-1}\boldsymbol{U}_s^\top \boldsymbol{\Sigma}_s \boldsymbol{\theta}_s^\star, \qquad \boldsymbol{Q} = \big[\boldsymbol{\theta}_1^\circ - \boldsymbol{\theta}_T^\star, \ldots, \boldsymbol{\theta}_S^\circ - \boldsymbol{\theta}_T^\star, \boldsymbol{b}_T\big],$$

where $\boldsymbol{b}_T = \boldsymbol{\theta}_T^\circ - \boldsymbol{\theta}_T^\star$ is the target representation bias. We work in the asymptotic limit of source and target sample sizes, while holding S, D and k fixed. Assume that $\overline{\boldsymbol{H}} := \boldsymbol{Q}^\top \boldsymbol{\Sigma}_T \boldsymbol{Q} \succ 0$. We estimate weights by minimizing two-fold target out-of-fold squared loss subject to $\boldsymbol{w} \geq 0$ and $\mathbf{1}^\top \boldsymbol{w} = 1$, then refit the target predictor on all target observations. The last coordinate of the weight vector corresponds to the target domain, with $w_{S+1} = w_T$. $\widehat{\boldsymbol{\theta}}_{S+1}$ denotes the target-only coefficient refitted on all target observations. Write $\widehat{\boldsymbol{\theta}}_{\text{AdaST}}$ for the aggregate and $\mathcal{R}_T$ for its squared prediction risk on an independent target observation. The full estimator and regularity assumptions are given in Appendix D.

Theorem 1 (Asymptotic risk and weight convergence) *Under the assumptions in Appendix D.1,*

$$\widehat{\boldsymbol{w}} \xrightarrow{a.s.} \boldsymbol{w}^\star = \arg\min_{\substack{\boldsymbol{w}\geq 0 \\ \mathbf{1}^\top \boldsymbol{w}=1}} \boldsymbol{w}^\top \overline{\boldsymbol{H}} \boldsymbol{w},$$

$$\mathcal{R}_T(\widehat{\boldsymbol{\theta}}_{\textit{AdaST}}) \xrightarrow{a.s.} \sigma_T^2 + (\boldsymbol{w}^\star)^\top \overline{\boldsymbol{H}} \boldsymbol{w}^\star \leq \sigma_T^2 + \overline{H}_{S+1,S+1}.$$

The limiting weights are unique. Equality with the target-only risk holds exactly when $\boldsymbol{w}^\star = \mathbf{e}_{S+1}$, the target-only weight vector.

Moreover, if $n_s \geq c n_T$ for all $s \in [S]$ and some fixed $c > 1$, then $\mathbb{E}|\widehat{w}_T - w_T^\star| = O(n_T^{-1/2})$.

The proof and finite-sample bounds are given in Appendices D and E. The asymptotic result shows how complementary source predictors can reduce target representation bias, yielding a limiting risk no greater than that of the target-only predictor. Under the stated sample-size conditions, the finite-sample rate quantifies how quickly the learned target weight approaches its population-optimal value and gives a vanishing bound on its expectation when sources alone are optimal.

3.2 Synthetic Validation

We study nonnegative least-squares aggregation with coefficient shifts $\boldsymbol{\Delta}_s = \boldsymbol{\theta}_s^\star - \boldsymbol{\theta}_T^\star = \tau \boldsymbol{d}_s$, where $\boldsymbol{d}_s \in \mathrm{col}(\boldsymbol{U}_s)$ and $\frac{1}{S}\sum_s \boldsymbol{d}_s^\top \boldsymbol{\Sigma}_T \boldsymbol{d}_s = 1$. The projected discrepancies satisfy $\widetilde{\boldsymbol{\Delta}}_s(\tau) = \boldsymbol{\theta}_s^\circ(\tau) - \boldsymbol{\theta}_T^\star = \boldsymbol{c}_s + \tau \boldsymbol{d}_s$, with $\boldsymbol{c}_s = \widetilde{\boldsymbol{\Delta}}_s(0)$. τ quantifies how far the source coefficients drift from the target coefficient, with smaller values corresponding to source coefficients closer to the target coefficient, while larger values indicate greater source–target heterogeneity. Even at $\tau = 0$, different prediction subspaces can produce different projected coefficients. We consider $\tau \in \{0.4, 0.5, 0.6, 0.75\}$ and $n_T \in \{100, 200, 400, 800, 1600, 3200\}$, with 200 replications per setting.

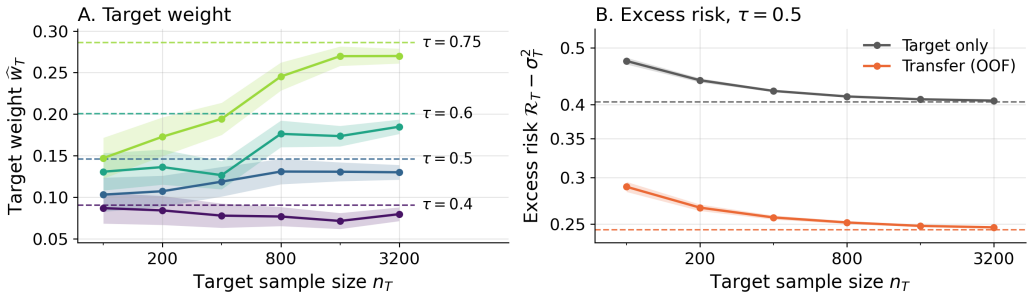


Figure 2: Synthetic validation of simplex-constrained least-squares aggregation. **A.** Target weight $\widehat{w}_T$ against n_T for several source drifts τ. **B.** Excess risk $\mathcal{R}_T - \sigma_T^2$ against n_T at $\tau = 0.5$. Solid lines summarize Monte Carlo means with 95% confidence bands for the means. Dashed lines mark the constrained population weights and risks from Theorem 1.

Figure 2 **A** shows that for larger τ, the learned target weight $\widehat{w}_T$ approaches its limit $w_T^\star(\tau)$ as the sample sizes grow, and this limit increases with τ. Intuitively, larger values of τ imply that the source domains are less informative for the target, so more weight shifts to the target. In this regime, the empirical target weight generally increases with n_T, as additional target observations improve the target fit and reduce the need to borrow information from mismatched source domains. Figure 2 **B** shows that the excess risks $\mathcal{R}_T - \sigma_T^2$ of the transfer and target-only estimators converge to their limits, with the transfer limit strictly below the target-only one.

4 Experiments

In this section, we evaluate AdaST primarily on cross-domain hallucination detection. We further examine its generalization to reasoning-step correctness. We also analyze the learned representations and evaluate activation steering based on the resulting directions.

4.1 Experimental Setup

Tasks and datasets. We consider several open-weight language models, including OPT-6.7B (Zhang et al., 2022), Llama-3.1-8B (Grattafiori et al., 2024), Gemma-7B (Gemma Team et al., 2024), Mistral-7B (Jiang et al., 2023), Qwen2-7B (Yang et al., 2024), and DeepSeek-LLM-7B (DeepSeek-AI et al., 2024). Across these models, we evaluate two behavioral prediction tasks, hallucination detection and reasoning-step correctness classification. Hallucination detection is evaluated on Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), CoQA (Reddy et al., 2019), HaluEval-QA (Li et al., 2023a), SimpleQA (Wei et al., 2024), and HotpotQA (Yang et al., 2018). Reasoning-step correctness is evaluated on PRM800K (Lightman et al., 2024), which contains questions from seven mathematical domains, including Algebra, Counting and Probability, Geometry, Intermediate Algebra, Number Theory, Prealgebra, and Precalculus.

Evaluation metrics. We use AUROC as the primary discrimination metric, defined as $\mathrm{AUROC} = \Pr(s(x^+) > s(x^-))$. Here, $s(\cdot)$ denotes the detector score and x^+ and x^- denote positive and negative responses. We also report Macro-F1, defined as $\text{Macro-F1} = \frac{1}{2}\left(F1_{\text{positive}} + F1_{\text{negative}}\right)$, where $F1_c = \frac{2P_cR_c}{P_c+R_c}$ for a class c. We report Macro-F1 rather than the F1 score of either class

because the response labels can be imbalanced. Macro-F1 gives equal weight to truthful and hallucinated responses and therefore avoids a high score driven primarily by the majority class.

Baselines. We compare against six hallucination detection baselines. HARP (Hu et al., 2026a) derives a reasoning subspace from the model's unembedding matrix and performs detection on hidden states projected onto this subspace. SinkProbe (Binkowski et al., 2026) detects hallucinations using statistics of attention-sink patterns derived from attention maps. LLMsKnow (Orgad et al., 2025) detects hallucinations from hidden states at specific answer-token positions. HaloScope (Du et al., 2024) uses a latent subspace to separate unlabeled generations into candidate truthful and hallucinated sets, which are then used to train a binary truthfulness classifier. PRISM-SAPLMA (Zhang et al., 2025b) applies PRISM prompts before extracting hidden states and feeds the resulting representations to SAPLMA for cross-domain detection. CrossHallu (Alansari et al., 2026) trains an MLP detector on source-domain features derived from hidden states, attention, and token probabilities, and studies its transfer to unseen target domains.

Implementation details. For hallucination detection, we use six QA datasets as six domains. Each dataset is treated in turn as the target domain, with the remaining five datasets serving as source domains. Each question is augmented with two semantically equivalent paraphrases, giving three question versions in total. We generate five responses per version and label each response by comparing it with all available reference answers using BLEURT-20 (Sellam et al., 2020; Pu et al., 2021). A response is labeled positive if its maximum BLEURT score exceeds 0.5 and negative otherwise. The target-label budget n_T counts original questions, with all paraphrases and generated responses from the same question kept together.

For reasoning-step correctness, we use PRM800K reasoning prefixes from the seven MATH subjects. In our transfer setting, each subject is treated in turn as the target, with the remaining six serving as sources. For each prefix, the evaluated model generates eight candidate next steps, which are scored by Qwen2.5-Math-PRM-7B (Zhang et al., 2025c). We retain the highest-scoring candidate with score at least 0.8 as positive and the lowest-scoring candidate with score at most 0.2 as negative, discarding prefixes for which both classes are unavailable. The target-label budget n_T counts retained reasoning prefixes.

For both tasks, representations and domain-specific detectors are constructed as described in Section 2, and transfer weights are estimated using three-fold cross-fitting. Model-specific layer selection are provided in Appendix B.1.

4.2 Main Results

Table 1: Performance of target-only detectors and AdaST across different target-label budgets, averaged over six datasets for hallucination detection and seven PRM800K MATH domains for reasoning-step correctness. Subscripts report standard errors.

		$n_T = 25$		$n_T = 50$		$n_T = 100$		Full	
Task	Data access	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Hallucination detection	Target-only	70.1 (1.5)	60.8 (1.4)	74.6 (1.3)	62.8 (1.3)	77.4 (1.3)	64.9 (1.3)	82.3 (0.9)	69.0 (1.1)
	AdaST	**77.9** (1.3)	**64.3** (1.3)	**79.2** (1.2)	**66.2** (1.2)	**80.6** (0.9)	**66.7** (1.2)	**83.0** (0.8)	**69.3** (1.1)
Reasoning-step correctness	Target-only	81.3 (0.9)	72.1 (0.9)	85.3 (0.8)	76.2 (0.7)	87.4 (0.8)	78.4 (0.9)	88.5 (0.8)	80.1 (0.8)
	AdaST	**89.0** (0.8)	**80.0** (0.9)	**89.3** (0.8)	**80.6** (0.9)	**89.9** (0.8)	**81.5** (0.9)	**89.7** (0.7)	**81.3** (0.9)

Table 1 compares the target-only detector with AdaST across the two tasks and different target-label budgets. AdaST consistently improves both AUROC and Macro-F1 across the two tasks, with larger gains when fewer target labels are available. With limited target data, the target detector is estimated with greater uncertainty. AdaST can compensate for this by assigning weight to source detectors that remain informative on held-out target examples. As more target labels become available, the target detector becomes more reliable and the additional benefit of source information decreases. Results by model and target dataset are reported in Appendices A.1 and A.2.

Table 2 compares AdaST with existing hallucination detectors at $n_T = 25$. AdaST improves over the domain-specific detector across both models and achieves the strongest average performance in

Table 2: Hallucination detection at $n_T = 25$ labeled target questions. The best average within each model is underlined. Subscript reports standard error

		NQ		TriviaQA		CoQA		HaluEval		SimpleQA		HotpotQA		Avg.	
Model	Method	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
OPT-6.7B	HaloScope	61.0 (3.7)	51.3 (2.0)	69.7 (1.3)	59.8 (0.8)	50.9 (3.4)	49.5 (1.5)	82.2 (1.6)	73.9 (0.9)	55.6 (7.7)	**55.4** (1.8)	59.6 (1.0)	52.1 (1.8)	63.2 (4.2)	57.0 (3.4)
	CrossHallu	78.1 (0.5)	58.6 (0.9)	80.5 (0.5)	46.0 (0.2)	**69.0** (0.4)	54.7 (1.1)	75.7 (0.4)	46.1 (1.1)	65.2 (0.5)	49.1 (0.2)	**80.3** (0.3)	54.6 (1.0)	74.8 (2.4)	51.5 (1.8)
	HARP	71.7 (1.4)	53.6 (1.2)	76.7 (1.0)	63.1 (0.9)	63.9 (1.2)	57.0 (1.8)	84.3 (0.5)	71.2 (3.4)	64.1 (3.7)	48.1 (2.7)	65.1 (1.5)	52.1 (0.9)	71.0 (3.1)	57.5 (3.2)
	PRISM-SAPLMA	57.3 (2.4)	49.2 (0.7)	68.9 (1.3)	50.6 (2.5)	55.3 (0.8)	49.5 (1.4)	69.6 (1.4)	56.0 (2.4)	62.2 (2.3)	48.9 (0.1)	60.1 (1.6)	48.9 (1.1)	62.2 (2.2)	50.5 (1.0)
	SinkProbe	70.5 (1.7)	55.1 (1.3)	78.4 (1.0)	64.0 (3.4)	60.6 (1.0)	49.6 (3.1)	83.6 (0.5)	62.6 (4.4)	**72.8** (1.5)	52.2 (1.1)	70.9 (1.2)	53.5 (1.3)	72.8 (2.9)	56.2 (2.7)
	Target-only (Ours)	71.1 (2.9)	54.0 (1.5)	76.0 (1.5)	64.5 (1.6)	63.9 (1.4)	59.1 (1.3)	85.1 (0.5)	**78.4** (0.6)	63.9 (6.9)	55.3 (2.2)	70.1 (2.2)	55.8 (1.3)	71.7 (3.0)	61.2 (3.5)
	AdaST (Ours)	**82.3** (0.5)	**58.8** (1.2)	**81.4** (0.7)	**66.7** (2.2)	66.0 (0.5)	**61.0** (0.5)	**86.7** (0.6)	78.3 (0.7)	72.1 (6.6)	55.0 (1.8)	74.4 (1.0)	**56.7** (0.8)	**77.1** (2.8)	**62.7** (3.3)
Mistral-7B	HaloScope	53.7 (3.0)	50.7 (1.2)	55.5 (6.4)	52.4 (1.5)	69.8 (9.0)	54.8 (6.0)	48.3 (6.3)	50.4 (2.1)	52.4 (3.1)	44.8 (1.3)	62.3 (5.1)	48.8 (6.0)	57.0 (2.9)	50.3 (1.2)
	CrossHallu	80.3 (0.1)	46.8 (0.0)	82.3 (0.2)	47.0 (0.1)	77.6 (0.3)	48.0 (0.1)	72.5 (0.5)	47.1 (0.0)	81.4 (0.2)	48.0 (0.0)	**84.1** (0.1)	48.0 (0.1)	79.7 (1.5)	47.5 (0.2)
	HARP	67.7 (1.1)	54.4 (1.2)	71.5 (0.8)	59.8 (0.8)	74.0 (0.7)	58.6 (0.9)	65.5 (0.6)	55.9 (0.5)	68.2 (0.8)	54.8 (0.7)	71.1 (1.1)	56.7 (0.8)	69.7 (1.1)	56.7 (0.8)
	PRISM-SAPLMA	62.9 (1.6)	51.2 (1.0)	66.8 (1.1)	53.3 (1.2)	69.9 (1.8)	51.8 (1.5)	64.6 (1.3)	51.3 (1.2)	69.6 (1.1)	54.0 (1.2)	65.9 (1.3)	50.5 (0.8)	66.6 (1.0)	52.0 (0.5)
	LLMsKnow	74.4 (0.7)	61.7 (0.6)	78.8 (0.6)	66.7 (0.4)	82.1 (0.3)	**63.1** (0.6)	77.4 (0.5)	64.2 (0.5)	77.1 (0.8)	61.2 (0.7)	77.7 (0.4)	60.3 (1.3)	77.9 (0.9)	62.9 (0.9)
	SinkProbe	75.0 (0.4)	64.3 (2.7)	**84.3** (0.5)	68.6 (2.6)	79.6 (0.7)	61.0 (1.9)	**79.8** (0.5)	60.4 (2.1)	**84.3** (0.8)	62.3 (2.3)	74.8 (0.5)	64.3 (2.6)	79.6 (0.9)	63.5 (1.5)
	Target-only (Ours)	71.4 (1.5)	60.3 (0.8)	75.6 (0.8)	64.3 (0.4)	77.7 (0.5)	58.4 (0.6)	71.3 (0.9)	59.3 (1.0)	70.5 (1.1)	59.1 (0.5)	66.3 (2.6)	57.3 (1.1)	72.1 (1.5)	59.8 (0.9)
	AdaST (Ours)	**80.7** (0.4)	**66.9** (0.3)	83.9 (0.3)	**70.5** (0.3)	**82.5** (0.4)	62.4 (0.9)	79.6 (0.5)	**65.9** (0.5)	83.0 (0.3)	**64.3** (0.9)	83.3 (0.2)	**64.8** (0.6)	**82.2** (0.6)	**65.8** (1.1)

both AUROC and Macro-F1. We also observe that some baselines remain competitive in AUROC but perform worse in Macro-F1. This gap shows that strong ranking performance does not always lead to balanced classification. AdaST remains strong on both metrics.

4.3 Further Analysis

Learned transfer path. Figure 3 **A** shows the learned expert allocation at $n_T = 25$. The weights differ across target datasets, suggesting that different targets benefit from different source detectors rather than from a common source combination. Figure 3 **B** shows how the overall allocation changes with target supervision. With few target questions, most weights are placed on the source detectors. As n_T grows, the target detector becomes more reliable, while the relative benefit of source information diminishes, leading AdaST to place more weight on the target detector.Figure 3 **C** shows the same overall trend across model families.

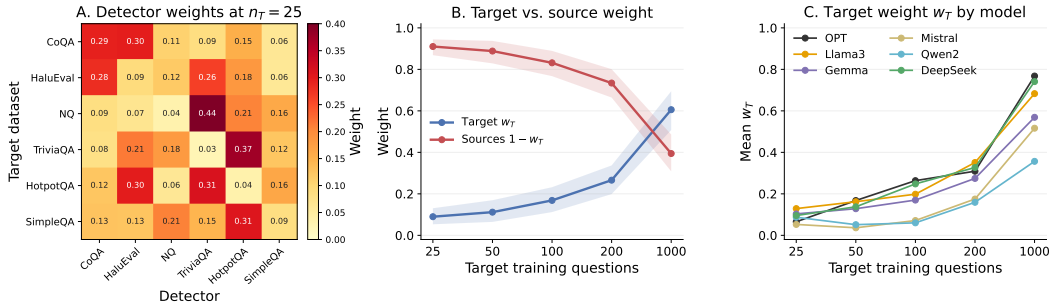


Figure 3: Learned AdaST combination weights. (a) Detector weights at $n_T = 25$. (b) Target weight w_T and source weight $1 - w_T$ across target training sizes. (c) Target weight w_T across model families.

Robustness to base detector choice. To test whether AdaST is robust to the choice of base detector, we use SinkProbe, the strongest single-dataset baseline in Table 2. For each domain, a SinkProbe expert is a logistic classifier fitted to its attention-sink features. The resulting classifier score is used as the domain-specific detector score in AdaST. Table 3 reports the results.

Table 3: Robustness to base detector choice at $n_T = 25$. AUROC (%).

SinkProbe as base detector	OPT-6.7B	Mistral-7B	Qwen2-7B
Target-only	73.7 (2.3)	79.6 (0.4)	70.9 (3.1)
AdaST	**77.8** (1.8)	**84.4** (0.4)	**79.3** (3.4)

Table 3 shows that AdaST improves SinkProbe across all three models, with an average gain of 5.77 AUROC points. SinkProbe relies on attention-sink features rather than the residual-stream representations used by our default detector. This suggests that the transfer gains are not specific to a particular detector construction.

Effect of paraphrase-aware estimation. Table 4 examines the effect of paraphrase augmentation. Within the evaluated range, adding paraphrased views improves AUROC for both the target-only detector and AdaST. The improvement suggests that paraphrases help produce a more stable domain-specific behavioral representation.

Table 4: Effect of paraphrase augmentation (AUROC, %).

Method	$P = 1$	$P = 2$	$P = 3$
Target-only	69.6 (4.0)	71.9 (2.9)	**72.8** (2.5)
AdaST	76.2 (2.8)	77.7 (2.3)	**78.2** (2.3)

From Detection to Intervention. We test whether the directions learned for detection can also steer generation. Using the residual-space directions in Equation 4 and the transfer weights in Equation 5, we define

$$\mathbf{u}_{\mathrm{mix}} = \frac{\sum_{s\in[S]\cup\{T\}} \widehat{w}_s \, \mathbf{v}_s/\kappa_s}{\left\| \sum_{s\in[S]\cup\{T\}} \widehat{w}_s \, \mathbf{v}_s/\kappa_s \right\|_2},$$

where κ_s is the target-score normalization factor for detector s. During generation, we steer the residual state at the detection layer as $\mathbf{h}_t^{(L)} \leftarrow \mathbf{h}_t^{(L)} + \alpha \mathbf{u}_{\mathrm{mix}}$, where $\alpha > 0$ controls the steering strength. The computation of κ_s and the full activation-steering implementation are detailed in Appendix B.2.

Table 5: Activation steering results on Llama-3.1-8B, averaged over three seeds.

Setting	Semantic accuracy	Hallucination	Abstention
No steering	27.2	67.8	5.0
AdaST steering	**30.0** (0.7)	**64.6** (0.9)	5.4 (0.4)

We use Phi-4 (Abdin et al., 2024) to assign semantic labels, from which we report semantic accuracy, hallucination and abstention. Abstention is the fraction of responses that explicitly decline to answer or provide no substantive answer. Steering improves semantic accuracy and reduces hallucination. The abstention rate remains largely unchanged, suggesting that the improvement is not simply driven by the model avoiding difficult questions.

5 Conclusion

We study adaptive transfer of LLM reliability signals across domains with AdaST. Experiments on hallucination detection and reasoning-step correctness show that transfer is most beneficial when target labels are limited, with the learned weights gradually shifting toward the target detector as more supervision becomes available. Our analysis characterizes this behavior in terms of source–target discrepancies and target representation bias. Future research will include extending linear classifiers to nonlinear ones.

AI Use Statement

Generative AI tools were used to assist with language editing, figure presentation, and feedback on the manuscript. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final content.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Aisha Alansari, Malak Alkhorasani, and Hamzah Luqman. CrossHallu: Do hallucination signals generalize across languages and domains in large language model's internals? *arXiv preprint arXiv:2607.04029*, 2026.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Guy Bar-Shalom, Fabrizio Frasca, Yaniv Galron, Yftah Ziser, and Haggai Maron. Beyond token probes: Hallucination detection via activation tensors with act-vit. In *Advances in Neural Information Processing Systems*, volume 38, pp. 95218–95248, 2025.

Jakub Binkowski, Kamil Adamczewski, and Tomasz Kajdanowicz. Attention sinks as internal signals for hallucination detection in large language models. In *International Conference on Machine Learning*, 2026.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

DeepSeek-AI et al. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Yongxin Deng, Zhen Fang, Yixuan Li, and Ling Chen. Beyond in-domain detection: Spikescore for cross-domain hallucination detection, 2026. URL https://arxiv.org/abs/2601.19245.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In *Advances in Neural Information Processing Systems*, 2024.

Gemma Team et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiang Guo, Darsh Shah, and Regina Barzilay. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4694–4703. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1498.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, 2024. doi: 10.1038/s41591-024-03097-1.

Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. HARP: Hallucination detection via reasoning subspace projection. In *International Conference on Learning Representations*, 2026a.

Yangfan Hu, Xuhan Tong, Haoyue Bai, Xi Ding, Shashank Muralidhar Bharadwaj, Siyang Cao, Robert Nowak, and Jiawei Zhang. Understanding why language models hallucinate: Testing reasoning against priors. *arXiv preprint arXiv:2607.00447*, 2026b.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857, 2022.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, volume 21, pp. 1041–1048, 2008.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 367–374, 2009.

Charles O'Neill, Slava Chalnev, Chi Chi Zhao, Max Kirkby, and Mudith Jayasekara. A single direction of truth: An observer model's linear residual probe exposes and steers contextual hallucinations. *arXiv preprint arXiv:2507.23221*, 2025.

OpenAI. The hugging face incident and the road ahead. OpenAI, August 2026. URL `https://openai.com/index/hugging-face-incident-and-the-road-ahead/`.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations*, 2025.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019. doi: 10.1109/ICCV.2019.00149.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning compact metrics for mt. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 751–762. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.58.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.704.

Nina I. Shamsi. Density ridge selective prediction for llm and vlm hallucination detection under calibration label scarcity, 2026. URL https://arxiv.org/abs/2606.10198.

Haoyi Song, Ruihan Ji, Naichen Shi, Fan Lai, and Raed Al Kontar. Inv-entropy: A fully probabilistic framework for uncertainty quantification in language models. *Advances in Neural Information Processing Systems*, 38:107889–107922, 2026.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Yujie Wu, Giovanni Parmigiani, and Boyu Ren. Multi-source domain adaptation for regression. *arXiv preprint arXiv:2312.05460*, 2023.

An Yang et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Boyang Zhang, Yicong Tan, Yun Shen, Ahmed Salem, Michael Backes, Savvas Zannettou, and Yang Zhang. Breaking agents: Compromising autonomous LLM agents through malfunction amplification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34964–34976, Suzhou, China, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1771. URL https://aclanthology.org/2025.emnlp-main.1771/.

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. Prompt-guided internal states for hallucination detection of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 21806–21818, 2025b.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025c.

Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. HaDeMiF: Hallucination detection and mitigation in large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VwOYxPScxB.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Additional Results

A.1 Per-Model Results

Table 6 breaks the $n_T = 25$ column of Table 1 down by model. The gain from the source domains is positive for every model on both tasks.

Table 6: Per-model breakdown of Table 1 at $n_T = 25$.

		OPT-6.7B		Llama-3.1-8B		Gemma-7B		Mistral-7B		Qwen2-7B		DeepSeek-7B		Avg.	
Task	Data access	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Hallucination	Target-only	71.7 (3.0)	61.2 (3.5)	73.0 (2.9)	65.0 (3.6)	72.7 (3.3)	64.0 (4.1)	72.1 (1.5)	59.8 (0.9)	56.3 (2.5)	51.7 (0.9)	74.5 (2.1)	63.0 (2.8)	70.1 (1.5)	60.8 (1.4)
	AdaST	**77.1** (2.9)	**62.7** (3.2)	**81.3** (2.1)	**69.2** (3.0)	**80.4** (2.5)	**66.4** (3.8)	**82.2** (0.6)	**65.8** (1.0)	**67.6** (3.8)	**57.3** (1.9)	**78.8** (2.4)	**64.5** (3.0)	**77.9** (1.3)	**64.3** (1.3)
Reasoning-step	Target-only	84.8 (1.5)	74.7 (1.5)	80.2 (2.0)	71.9 (2.5)	73.3 (1.8)	65.1 (1.6)	83.5 (1.2)	73.7 (1.5)	82.5 (1.7)	72.9 (2.1)	83.7 (1.2)	74.0 (1.3)	81.3 (0.9)	72.1 (0.9)
	AdaST	**92.3** (0.8)	**84.8** (0.6)	**87.5** (1.3)	**78.3** (1.3)	**82.5** (1.9)	**72.7** (1.8)	**92.4** (0.5)	**83.4** (0.8)	**86.5** (1.5)	**76.1** (1.1)	**92.8** (1.0)	**84.9** (1.4)	**89.0** (0.8)	**80.0** (0.9)

A.2 Per-Dataset Results

Tables 7 and 8 break Table 1 down by target dataset. The gain from the source domains is positive on every target dataset of both tasks at $n_T = 25$.

Table 7: AdaST for hallucination detection against the same detector without source data, by dataset and target label budget n_T.

		NQ		TriviaQA		CoQA		HaluEval		SimpleQA		HotpotQA		Avg.	
n_T	Data access	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
25	Target-only	68.2 (3.7)	57.2 (1.8)	76.0 (3.1)	67.3 (3.2)	68.0 (2.4)	60.0 (1.8)	78.2 (2.9)	70.2 (3.9)	63.9 (2.7)	54.2 (1.3)	65.9 (2.8)	55.6 (1.2)	70.1 (1.5)	60.8 (1.4)
	AdaST	**79.0** (4.4)	**62.2** (2.4)	**83.0** (1.2)	**71.2** (2.3)	**71.8** (2.2)	**62.4** (0.8)	**83.3** (1.1)	**73.3** (2.9)	**75.5** (3.1)	**56.3** (1.8)	**74.9** (3.1)	**60.6** (1.4)	**77.9** (1.3)	**64.3** (1.3)
50	Target-only	74.2 (3.8)	59.1 (1.6)	79.5 (2.0)	69.3 (2.8)	70.7 (2.1)	62.4 (1.6)	80.8 (1.9)	72.0 (3.2)	72.0 (3.6)	56.6 (0.9)	70.4 (2.7)	57.3 (1.2)	74.6 (1.4)	62.8 (1.3)
	AdaST	**80.3** (3.8)	**65.4** (2.0)	**83.3** (1.2)	**72.8** (2.2)	**73.5** (2.0)	**64.6** (1.1)	**83.9** (1.0)	**74.5** (2.6)	**76.3** (4.1)	**58.7** (1.4)	**77.7** (1.5)	**61.2** (1.3)	**79.2** (1.2)	**66.2** (1.2)
100	Target-only	78.3 (2.7)	61.8 (2.3)	81.7 (1.6)	71.4 (2.7)	73.3 (1.9)	64.1 (1.5)	83.4 (1.3)	74.3 (2.7)	74.5 (4.2)	58.7 (1.3)	73.4 (2.5)	59.4 (1.3)	77.4 (1.2)	64.9 (1.3)
	AdaST	**83.9** (1.2)	**65.3** (1.8)	**83.8** (1.0)	**72.8** (2.1)	**74.6** (1.8)	**65.4** (1.1)	**84.6** (1.0)	**75.4** (2.5)	**77.9** (3.2)	**58.9** (1.6)	**78.7** (1.4)	**62.2** (1.2)	**80.6** (1.0)	**66.7** (1.2)
Full	Target-only	81.9 (3.1)	66.0 (1.8)	85.2 (1.1)	74.8 (2.0)	78.0 (1.5)	68.2 (1.6)	86.6 (0.8)	77.0 (2.2)	**80.8** (2.7)	63.0 (1.6)	81.3 (0.7)	**65.1** (1.1)	82.3 (0.9)	69.0 (1.1)
	AdaST	**86.0** (1.0)	**66.9** (1.6)	**85.5** (1.0)	**74.9** (2.0)	**78.0** (1.4)	**68.3** (1.5)	**86.6** (0.8)	**77.2** (2.1)	80.7 (2.7)	**63.2** (1.5)	**81.3** (0.9)	65.1 (0.9)	**83.0** (0.8)	**69.3** (1.1)

Table 8: AdaST for Reasoning-step correctness against the same detector without source data, by MATH domain and target label budget n_T. Prealgebra has fewer than 100 labeled training questions, so the $n_T = 100$ row omits it and averages over the remaining six domains.

		Algebra		C.&Prob.		Geometry		Int. Alg.		N. Theory		Prealg.		Precalc.		Avg.	
n_T	Data access	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
25	Target-only	76.5 (1.4)	67.0 (2.3)	80.4 (1.2)	72.5 (1.9)	82.8 (0.7)	73.9 (0.7)	83.2 (1.0)	73.9 (1.3)	81.7 (1.0)	71.6 (1.6)	81.8 (1.3)	73.6 (1.7)	82.8 (1.1)	71.9 (1.8)	81.3 (0.9)	72.1 (0.9)
	AdaST	**84.5** (1.4)	**76.1** (2.2)	**89.4** (0.7)	**80.5** (1.1)	**90.8** (0.6)	**81.1** (1.3)	**90.2** (0.8)	**80.4** (1.6)	**90.5** (0.6)	**81.1** (1.5)	**86.9** (1.3)	**78.7** (1.8)	**90.9** (0.6)	**82.3** (0.9)	**89.0** (0.8)	**80.0** (0.9)
50	Target-only	80.8 (1.3)	72.9 (1.8)	85.6 (0.9)	77.2 (1.1)	86.9 (0.7)	76.4 (1.0)	86.0 (1.0)	77.4 (1.4)	86.8 (0.9)	77.9 (1.4)	83.7 (1.2)	74.1 (1.4)	87.4 (0.8)	77.9 (1.2)	85.3 (0.8)	76.2 (0.7)
	AdaST	**85.1** (1.4)	**76.8** (2.2)	**89.5** (0.6)	**80.4** (1.1)	**91.3** (0.7)	**82.6** (1.5)	**90.9** (0.7)	**81.6** (1.3)	**90.4** (0.7)	**82.5** (1.2)	**87.1** (1.3)	**78.8** (2.2)	**91.1** (0.7)	**81.8** (1.1)	**89.3** (0.8)	**80.6** (0.9)
100	Target-only	83.1 (1.3)	74.6 (1.9)	88.1 (0.8)	78.3 (1.3)	87.8 (0.7)	77.8 (1.5)	88.2 (0.7)	79.1 (1.3)	87.9 (0.9)	79.8 (1.3)	–	–	89.4 (0.8)	80.7 (1.3)	87.4 (0.8)	78.4 (0.9)
	AdaST	**85.7** (1.4)	**77.5** (2.1)	**89.7** (0.6)	**81.5** (1.3)	**91.3** (0.7)	**83.0** (1.4)	**90.8** (0.7)	**82.6** (1.3)	**90.6** (0.7)	**82.9** (1.3)	–	–	**91.4** (0.6)	**81.6** (1.4)	**89.9** (0.8)	**81.5** (1.0)
Full	Target-only	84.9 (1.3)	76.6 (1.8)	89.2 (0.7)	81.2 (1.0)	89.5 (0.7)	82.1 (1.2)	90.8 (0.7)	**82.6** (1.3)	88.8 (0.9)	80.6 (1.3)	85.7 (1.1)	76.7 (1.6)	90.6 (0.8)	80.5 (1.2)	88.5 (0.8)	80.1 (0.8)
	AdaST	**86.3** (1.3)	**78.7** (1.9)	**90.1** (0.6)	**81.3** (1.2)	**91.3** (0.7)	**83.1** (1.4)	**91.0** (0.7)	82.3 (1.4)	**90.5** (0.6)	**82.7** (1.7)	**87.3** (1.2)	**79.1** (1.9)	**91.4** (0.6)	**82.1** (1.2)	**89.7** (0.7)	**81.3** (0.9)

B Additional Implementation Details

B.1 Layer Selection

For each model and target domain T, we select the representation layer using labeled source data only. For each source domain $s \in [S]$, we divide its questions (or reasoning prefixes) into a 70% fitting partition and a 30% validation partition. At each candidate layer j, we estimate a mean-difference direction between the positive and negative classes using the fitting partition and evaluate its AUROC on the corresponding validation partition. Here, the positive class corresponds to truthful responses for hallucination detection and correct next steps for reasoning-step correctness. We select

$$L_{m,T} = \arg\max_j \frac{1}{S} \sum_{s \in [S]} \mathrm{AUROC}^{\mathrm{val}}_{m,s,j},$$

where m indexes the evaluated language model. The selected layer is then fixed for all target-label budgets, cross-fitting folds, and experimental repetitions for that model–target pair.

For hallucination detection, the five remaining QA datasets serve as source domains when one dataset is held out as the target. For reasoning-step correctness, the six remaining MATH subjects serve as source domains. Neither labeled target data nor the target test set is used for layer selection, preserving the target-label budget and avoiding target-test leakage. After selecting $L_{m,T}$, the source detectors are refitted using all available labeled data from their respective source domains.

B.2 Intervention Direction and Activation Steering

For detector $s \in [S] \cup \{T\}$, recall that its raw residual-space score is $\mathbf{v}_s^\top \mathbf{h}$. We denote it by $g_s(\mathbf{h})$.

To place the scores of different detectors on a common scale, we standardize them using the labeled target responses. Specifically, we define the mean raw score of detector s on the target set as

$$\bar{g}_s = \frac{1}{n_T PR} \sum_{q \in \mathcal{Q}_T} \sum_{p=1}^{P} \sum_{r=1}^{R} \mathbf{v}_s^\top \mathbf{h}_{qpr},$$

and define

$$\kappa_s = \sqrt{\frac{1}{n_T PR - 1} \sum_{q \in \mathcal{Q}_T} \sum_{p=1}^{P} \sum_{r=1}^{R} \left(\mathbf{v}_s^\top \mathbf{h}_{qpr} - \bar{g}_s\right)^2}.$$

Thus, κ_s is the empirical standard deviation of detector s's raw scores on the labeled target responses. The standardized detector score is therefore

$$\psi_s(\mathbf{h}) = \frac{\mathbf{v}_s^\top \mathbf{h} - \bar{g}_s}{\kappa_s}.$$

Since $\bar{g}_s$ is constant with respect to $\mathbf{h}$, the direction in residual space along which the standardized detector score increases is

$$\nabla_{\mathbf{h}} \psi_s(\mathbf{h}) = \frac{\mathbf{v}_s}{\kappa_s}.$$

The final AdaST detector combines the standardized detector scores as

$$\widehat{f}(\mathbf{h}) = \widehat{b} + \sum_{s \in [S] \cup \{T\}} \widehat{w}_s \psi_s(\mathbf{h}).$$

Its corresponding residual-space direction is therefore

$$\nabla_{\mathbf{h}} \widehat{f}(\mathbf{h}) = \sum_{s \in [S] \cup \{T\}} \widehat{w}_s \frac{\mathbf{v}_s}{\kappa_s}.$$

We normalize this vector to unit norm to obtain the steering direction $\mathbf{u}_{\mathrm{mix}}$ used in the intervention experiment. We next describe how the resulting direction is applied during generation. We implement activation steering using a forward hook on transformer block L. During the initial prompt pass, the hook adds $\alpha \mathbf{u}_{\mathrm{mix}}$ only to the block output at the final prompt position. During cached autoregressive decoding, the same update is applied to the current token position at every decoding step:

$$\mathbf{h}_t^{(L)} \leftarrow \mathbf{h}_t^{(L)} + \alpha \mathbf{u}_{\mathrm{mix}}.$$

The modified state is then passed to block $L + 1$ to predict the next token. The steering direction $\mathbf{u}_{\mathrm{mix}}$ and strength α are fixed throughout the generation of a response. No earlier prompt positions or other transformer layers are modified.

For a unit steering direction $\mathbf{u}_{\mathrm{mix}}$, we first compute its activation scale

$$A_{\mathrm{mix}} = \max_{i \in \mathcal{V}} \left| \left\langle \mathbf{h}_{i,-1}^{(L)}, \mathbf{u}_{\mathrm{mix}} \right\rangle \right|,$$

where $\mathcal{V}$ denotes the steering validation set and $\mathbf{h}_{i,-1}^{(L)}$ is the residual state of the final prompt token at layer L. We evaluate strengths of the form $\alpha = cA_{\mathrm{mix}}$ for $c \in \{0, 0.1, 0.3, 1, 3, 10, 20, 40\}$ and select the value that maximizes mean BLEURT on the validation responses. Ties are resolved in favor of the smaller magnitude. The selected α is then frozen and evaluated on 500 disjoint test questions.

C Additional Lemmas

Lemma 1 (Sandwiched sample second moments.) *Let $x_1, \dots, x_n \in \mathbb{R}^D$ be i.i.d. with $\mathbb{E}\|x_\ell\|_2^2 < \infty$ and $\mathbb{E}[x_\ell x_\ell^\top] = \Sigma$, and let $\widehat{\Sigma} = \frac{1}{n}\sum_{\ell=1}^n x_\ell x_\ell^\top$. Let $A \in \mathbb{R}^{D\times m_A}$ and $B \in \mathbb{R}^{D\times m_B}$ be fixed matrices whose numbers of columns do not depend on n. With D, A, and B fixed, as $n \to \infty$,*

$$A^\top \widehat{\Sigma} B \xrightarrow{\text{a.s.}} A^\top \Sigma B.$$

If in addition $A^\top \Sigma A$ is positive definite, then almost surely $A^\top \widehat{\Sigma} A$ is invertible for all large n and $(A^\top \widehat{\Sigma} A)^{-1} \to (A^\top \Sigma A)^{-1}$.

Proof. Fix an entry (a, b) with $a \le m_A$ and $b \le m_B$, and let e_a, e_b denote standard basis vectors of $\mathbb{R}^{m_A}$ and $\mathbb{R}^{m_B}$. The entry is a sample mean,

$$\big(A^\top \widehat{\Sigma} B\big)_{ab} = \frac{1}{n}\sum_{\ell=1}^n Y_\ell, \qquad Y_\ell = \big(e_a^\top A^\top x_\ell\big)\big(e_b^\top B^\top x_\ell\big).$$

The variables Y_ℓ are i.i.d. across ℓ, being a fixed function of the i.i.d. vectors x_ℓ. They are integrable by the Cauchy Schwarz inequality,

$$\mathbb{E}|Y_\ell| \le \big(\mathbb{E}(e_a^\top A^\top x_\ell)^2\big)^{1/2}\big(\mathbb{E}(e_b^\top B^\top x_\ell)^2\big)^{1/2} = \big(e_a^\top A^\top \Sigma A e_a\big)^{1/2}\big(e_b^\top B^\top \Sigma B e_b\big)^{1/2} < \infty.$$

Their mean is $\mathbb{E}Y_\ell = (A^\top \Sigma B)_{ab}$. Kolmogorov's strong law of large numbers gives $\frac{1}{n}\sum_\ell Y_\ell \to (A^\top \Sigma B)_{ab}$ almost surely. This holds on an event Ω_{ab} with $\mathbb{P}(\Omega_{ab}) = 1$. There are $m_A m_B$ entries, a finite number that does not depend on n, so $\mathbb{P}(\bigcap_{a,b}\Omega_{ab}) = 1$. On this event every entry converges. Entrywise convergence of matrices of fixed size is convergence in every matrix norm, which proves the first claim.

For the second claim, work on the same probability one event with $A = B$. By Weyl's inequality,

$$\lambda_{\min}(A^\top \widehat{\Sigma} A) \ge \lambda_{\min}(A^\top \Sigma A) - \|A^\top \widehat{\Sigma} A - A^\top \Sigma A\|_{\text{op}}.$$

The right side converges to $\lambda_{\min}(A^\top \Sigma A) > 0$, so $A^\top \widehat{\Sigma} A$ is invertible for all large n. Continuity of matrix inversion gives $(A^\top \widehat{\Sigma} A)^{-1} \to (A^\top \Sigma A)^{-1}$.

Lemma 2 (Projected noise sample means.) *Let $(x_\ell, \varepsilon_\ell)$, $\ell = 1, \dots, n$, be i.i.d. pairs with $\mathbb{E}[\varepsilon_\ell \mid x_\ell] = 0$, $\mathbb{E}\varepsilon_\ell^2 < \infty$, and $\mathbb{E}\|x_\ell\|_2^2 < \infty$. Write $\Sigma = \mathbb{E}[x_\ell x_\ell^\top]$, $X = [x_1, \dots, x_n] \in \mathbb{R}^{D\times n}$, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$. Let $A \in \mathbb{R}^{D\times m_A}$ be a fixed matrix whose number of columns does not depend on n. With D and A fixed, as $n \to \infty$,*

$$\frac{1}{n} A^\top X \varepsilon = \frac{1}{n}\sum_{\ell=1}^n (A^\top x_\ell)\varepsilon_\ell \xrightarrow{\text{a.s.}} 0.$$

Proof. Fix a coordinate $a \le m_A$ and let e_a be the ath standard basis vector of $\mathbb{R}^{m_A}$. The coordinate is the sample mean $\frac{1}{n}\sum_\ell Y_\ell$ with $Y_\ell = (e_a^\top A^\top x_\ell)\varepsilon_\ell$. These variables are i.i.d. across ℓ and integrable by the Cauchy Schwarz inequality,

$$\mathbb{E}|Y_\ell| \le \big(\mathbb{E}(e_a^\top A^\top x_\ell)^2\big)^{1/2}\big(\mathbb{E}\varepsilon_\ell^2\big)^{1/2} = \big(e_a^\top A^\top \Sigma A e_a\big)^{1/2}\big(\mathbb{E}\varepsilon_\ell^2\big)^{1/2} < \infty.$$

Their mean is zero by the tower property,

$$\mathbb{E}Y_\ell = \mathbb{E}\Big[(e_a^\top A^\top x_\ell)\mathbb{E}[\varepsilon_\ell \mid x_\ell]\Big] = 0.$$

Kolmogorov's strong law of large numbers gives $\frac{1}{n}\sum_\ell Y_\ell \to 0$ almost surely. Since there are finitely many coordinates, the entire vector converges to zero almost surely.

D Proof of Theorem 1

D.1 Model and Assumptions

We use the Gaussian linear model and fixed prediction subspaces defined in the main text. Observations are i.i.d. within each domain, and datasets from different domains are independent. Noise is independent of features, with mean zero and fixed finite variance. The target test observation is independent of all training data.

All source and target sample sizes tend to infinity, while S, D, k, the prediction subspaces, and the population distributions remain fixed. Assume that the projected covariances $\boldsymbol{U}_s^\top \boldsymbol{\Sigma}_s \boldsymbol{U}_s$ are positive definite in every domain and that $\overline{\boldsymbol{H}} = \boldsymbol{Q}^\top \boldsymbol{\Sigma}_T \boldsymbol{Q} \succ 0$. The target fold assignments are fixed across the asymptotic sequence and independent of the observations, with $n_{T,f}/n_T \to \rho_f \in (0,1)$ for $f = 1, 2$.

D.2 Two-Fold Out-of-Fold Weight Estimation

Write $\boldsymbol{X}_s = [\boldsymbol{x}_{s,1}, \ldots, \boldsymbol{x}_{s,n_s}]$, $\boldsymbol{y}_s = (y_{s,1}, \ldots, y_{s,n_s})^\top$, and $\boldsymbol{\varepsilon}_s = (\varepsilon_{s,1}, \ldots, \varepsilon_{s,n_s})^\top$. For each source domain, fit the projected linear regression

$$\widehat{\boldsymbol{\beta}}_s = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \frac{1}{n_s} \|\boldsymbol{y}_s - \boldsymbol{X}_s^\top \boldsymbol{U}_s \boldsymbol{\beta}\|_2^2, \qquad \widehat{\boldsymbol{\theta}}_s = \boldsymbol{U}_s \widehat{\boldsymbol{\beta}}_s.$$

Split the target observations into two disjoint folds with sample sizes $n_{T,1} + n_{T,2} = n_T$. Using only fold $f \in \{1, 2\}$, fit the projected linear regression in the target subspace

$$\widehat{\boldsymbol{\beta}}_{T,f} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \frac{1}{n_{T,f}} \|\boldsymbol{y}_{T,f} - \boldsymbol{X}_{T,f}^\top \boldsymbol{U}_T \boldsymbol{\beta}\|_2^2, \qquad \widehat{\boldsymbol{\theta}}_{T,f} = \boldsymbol{U}_T \widehat{\boldsymbol{\beta}}_{T,f}.$$

The aggregation weights $\boldsymbol{w} = (w_1, \ldots, w_S, w_{S+1})^\top$ are estimated by

$$\widehat{\boldsymbol{w}} = \arg \min_{\substack{\boldsymbol{w} \geq 0 \\ \mathbf{1}^\top \boldsymbol{w} = 1}} \frac{1}{n_T} \left[\left\| \boldsymbol{y}_{T,1} - \boldsymbol{X}_{T,1}^\top \left(\sum_{s=1}^S w_s \widehat{\boldsymbol{\theta}}_s + w_{S+1} \widehat{\boldsymbol{\theta}}_{T,2} \right) \right\|_2^2 + \left\| \boldsymbol{y}_{T,2} - \boldsymbol{X}_{T,2}^\top \left(\sum_{s=1}^S w_s \widehat{\boldsymbol{\theta}}_s + w_{S+1} \widehat{\boldsymbol{\theta}}_{T,1} \right) \right\|_2^2 \right].$$

Any measurable minimizer can be used if the empirical optimum is not unique.

D.3 Risk Calculation

The projected linear regression using the full target dataset is

$$\widehat{\boldsymbol{\beta}}_T = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \frac{1}{n_T} \|\boldsymbol{y}_T - \boldsymbol{X}_T^\top \boldsymbol{U}_T \boldsymbol{\beta}\|_2^2, \qquad \widehat{\boldsymbol{\theta}}_{S+1} = \boldsymbol{U}_T \widehat{\boldsymbol{\beta}}_T.$$

Hence, the transferred estimator is

$$\widehat{\boldsymbol{\theta}}_{\text{AdaST}} = \sum_{s=1}^S \widehat{w}_s \widehat{\boldsymbol{\theta}}_s + \widehat{w}_{S+1} \widehat{\boldsymbol{\theta}}_{S+1}, \tag{7}$$

with the last coordinate of the weight vector corresponds to the target domain and $\widehat{w}_{S+1} = \widehat{w}_T$. For an independent target observation $(\boldsymbol{x}_0, y_0)$, the noise has mean zero and is independent of the features, so

$$\mathcal{R}_T(\widehat{\boldsymbol{\theta}}_{\text{AdaST}}) = \mathbb{E}\big[(y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\theta}}_{\text{AdaST}})^2 \mid \widehat{\boldsymbol{\theta}}_{\text{AdaST}}\big] = \sigma_T^2 + (\widehat{\boldsymbol{\theta}}_{\text{AdaST}} - \boldsymbol{\theta}_T^\star)^\top \boldsymbol{\Sigma}_T (\widehat{\boldsymbol{\theta}}_{\text{AdaST}} - \boldsymbol{\theta}_T^\star).$$

Using Equation 7 and the constraint $\mathbf{1}^\top \widehat{\boldsymbol{w}} = 1$,

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{\text{AdaST}} - \boldsymbol{\theta}_T^\star &= \sum_{s=1}^S \widehat{w}_s \widehat{\boldsymbol{\theta}}_s + \widehat{w}_{S+1} \widehat{\boldsymbol{\theta}}_{S+1} - \left(\sum_{s=1}^S \widehat{w}_s + \widehat{w}_{S+1} \right) \boldsymbol{\theta}_T^\star \\ &= \sum_{s=1}^S \widehat{w}_s (\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_T^\star) + \widehat{w}_{S+1} (\widehat{\boldsymbol{\theta}}_{S+1} - \boldsymbol{\theta}_T^\star). \end{aligned}$$

Here, $\widehat{\boldsymbol{\theta}}_{S+1}$ is the target-only coefficient. For each source task, the source–target shift, representation bias, and estimation error are

$$\boldsymbol{\Delta}_s = \boldsymbol{\theta}_s^\star - \boldsymbol{\theta}_T^\star, \qquad \boldsymbol{b}_s = \boldsymbol{\theta}_s^\circ - \boldsymbol{\theta}_s^\star, \qquad \boldsymbol{\xi}_s = \widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s^\circ.$$

The effective discrepancy between the projected source coefficient and the target coefficient is therefore

$$\widetilde{\boldsymbol{\Delta}}_s = \boldsymbol{\theta}_s^\circ - \boldsymbol{\theta}_T^\star = \boldsymbol{\Delta}_s + \boldsymbol{b}_s, \qquad \widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_T^\star = \widetilde{\boldsymbol{\Delta}}_s + \boldsymbol{\xi}_s.$$

For the target task, the representation bias and estimation error are

$$\boldsymbol{b}_T = \boldsymbol{\theta}_T^\circ - \boldsymbol{\theta}_T^\star, \qquad \boldsymbol{\xi}_{S+1} = \widehat{\boldsymbol{\theta}}_{S+1} - \boldsymbol{\theta}_T^\circ.$$

Hence,

$$\widehat{\boldsymbol{\theta}}_{\text{AdaST}} - \boldsymbol{\theta}_T^\star = \sum_{s=1}^{S} \widehat{w}_s(\widetilde{\boldsymbol{\Delta}}_s + \boldsymbol{\xi}_s) + \widehat{w}_{S+1}(\boldsymbol{b}_T + \boldsymbol{\xi}_{S+1}),$$

and expanding the quadratic risk gives

$$\begin{aligned}\mathcal{R}_T(\widehat{\boldsymbol{\theta}}_{\text{AdaST}}) = \sigma_T^2 &+ \sum_{s=1}^{S}\sum_{s'=1}^{S} \widehat{w}_s\widehat{w}_{s'}\Big[\widetilde{\boldsymbol{\Delta}}_s^\top \boldsymbol{\Sigma}_T \widetilde{\boldsymbol{\Delta}}_{s'} + \widetilde{\boldsymbol{\Delta}}_s^\top \boldsymbol{\Sigma}_T \boldsymbol{\xi}_{s'} + \boldsymbol{\xi}_s^\top \boldsymbol{\Sigma}_T \widetilde{\boldsymbol{\Delta}}_{s'} + \boldsymbol{\xi}_s^\top \boldsymbol{\Sigma}_T \boldsymbol{\xi}_{s'}\Big] \\ &+ 2\sum_{s=1}^{S} \widehat{w}_s\widehat{w}_{S+1}\Big[\widetilde{\boldsymbol{\Delta}}_s^\top \boldsymbol{\Sigma}_T \boldsymbol{b}_T + \widetilde{\boldsymbol{\Delta}}_s^\top \boldsymbol{\Sigma}_T \boldsymbol{\xi}_{S+1} + \boldsymbol{\xi}_s^\top \boldsymbol{\Sigma}_T \boldsymbol{b}_T + \boldsymbol{\xi}_s^\top \boldsymbol{\Sigma}_T \boldsymbol{\xi}_{S+1}\Big] \\ &+ \widehat{w}_{S+1}^2\Big[\boldsymbol{b}_T^\top \boldsymbol{\Sigma}_T \boldsymbol{b}_T + \boldsymbol{b}_T^\top \boldsymbol{\Sigma}_T \boldsymbol{\xi}_{S+1} + \boldsymbol{\xi}_{S+1}^\top \boldsymbol{\Sigma}_T \boldsymbol{b}_T + \boldsymbol{\xi}_{S+1}^\top \boldsymbol{\Sigma}_T \boldsymbol{\xi}_{S+1}\Big].\end{aligned}$$

We first show that the estimation errors vanish. Let $\widehat{\boldsymbol{\Sigma}}_s = \boldsymbol{X}_s\boldsymbol{X}_s^\top/n_s$. Substituting $\boldsymbol{y}_s = \boldsymbol{X}_s^\top\boldsymbol{\theta}_s^\star + \boldsymbol{\varepsilon}_s$ into the projected least-squares normal equations gives

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_s &= (\boldsymbol{U}_s^\top \widehat{\boldsymbol{\Sigma}}_s \boldsymbol{U}_s)^{-1} \boldsymbol{U}_s^\top \widehat{\boldsymbol{\Sigma}}_s \boldsymbol{\theta}_s^\star \\ &\quad + (\boldsymbol{U}_s^\top \widehat{\boldsymbol{\Sigma}}_s \boldsymbol{U}_s)^{-1} \frac{1}{n_s} \boldsymbol{U}_s^\top \boldsymbol{X}_s \boldsymbol{\varepsilon}_s.\end{aligned}$$

By Lemma 1,

$$\boldsymbol{U}_s^\top \widehat{\boldsymbol{\Sigma}}_s \boldsymbol{U}_s \xrightarrow{a.s.} \boldsymbol{U}_s^\top \boldsymbol{\Sigma}_s \boldsymbol{U}_s, \qquad \boldsymbol{U}_s^\top \widehat{\boldsymbol{\Sigma}}_s \boldsymbol{\theta}_s^\star \xrightarrow{a.s.} \boldsymbol{U}_s^\top \boldsymbol{\Sigma}_s \boldsymbol{\theta}_s^\star.$$

The first limit is positive definite, so the sample matrix is almost surely invertible for all sufficiently large sample sizes and its inverse converges to $(\boldsymbol{U}_s^\top \boldsymbol{\Sigma}_s \boldsymbol{U}_s)^{-1}$. By Lemma 2,

$$\frac{1}{n_s} \boldsymbol{U}_s^\top \boldsymbol{X}_s \boldsymbol{\varepsilon}_s \xrightarrow{a.s.} 0.$$

Therefore,

$$\widehat{\boldsymbol{\beta}}_s \xrightarrow{a.s.} (\boldsymbol{U}_s^\top \boldsymbol{\Sigma}_s \boldsymbol{U}_s)^{-1} \boldsymbol{U}_s^\top \boldsymbol{\Sigma}_s \boldsymbol{\theta}_s^\star = \boldsymbol{\beta}_s^\circ,$$

and hence

$$\boldsymbol{\xi}_s = \widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s^\circ = \boldsymbol{U}_s(\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^\circ) \xrightarrow{a.s.} 0.$$

The same argument with $s = T$ defines $\boldsymbol{\beta}_T^\circ$ and applies to the two target-fold estimators and the full-sample target estimator. Defining $\boldsymbol{\xi}_{T,f} = \widehat{\boldsymbol{\theta}}_{T,f} - \boldsymbol{\theta}_T^\circ$ for $f = 1, 2$, we obtain

$$\boldsymbol{\xi}_{T,1} \xrightarrow{a.s.} 0, \qquad \boldsymbol{\xi}_{T,2} \xrightarrow{a.s.} 0, \qquad \boldsymbol{\xi}_{S+1} \xrightarrow{a.s.} 0.$$

D.4 Asymptotic Out-of-Fold Weights

For fold $f \in \{1, 2\}$, substituting the target model and using $\mathbf{1}^\top \boldsymbol{w} = 1$, the residual in the out-of-fold objective becomes

$$\boldsymbol{\varepsilon}_{T,f} - \boldsymbol{X}_{T,f}^\top \left(\sum_{s=1}^{S} w_s(\widetilde{\boldsymbol{\Delta}}_s + \boldsymbol{\xi}_s) + w_{S+1}(\boldsymbol{b}_T + \boldsymbol{\xi}_{T,3-f}) \right).$$

Define

$$\widehat{\boldsymbol{E}}_f = \left[\widetilde{\boldsymbol{\Delta}}_1 + \boldsymbol{\xi}_1 \quad \cdots \quad \widetilde{\boldsymbol{\Delta}}_S + \boldsymbol{\xi}_S \quad \boldsymbol{b}_T + \boldsymbol{\xi}_{T,3-f}\right].$$

Expanding the squared residuals, the out-of-fold objective is

$$L(\boldsymbol{w}) = \frac{1}{n_T}\|\boldsymbol{\varepsilon}_T\|_2^2 - 2\boldsymbol{w}^\top \widehat{\boldsymbol{g}} + \boldsymbol{w}^\top \widehat{\boldsymbol{H}}\boldsymbol{w},$$

where

$$\widehat{\boldsymbol{g}} = \frac{1}{n_T}\sum_{f=1}^{2} \widehat{\boldsymbol{E}}_f^\top \boldsymbol{X}_{T,f}\boldsymbol{\varepsilon}_{T,f}, \qquad \widehat{\boldsymbol{H}} = \frac{1}{n_T}\sum_{f=1}^{2} \widehat{\boldsymbol{E}}_f^\top \boldsymbol{X}_{T,f}\boldsymbol{X}_{T,f}^\top \widehat{\boldsymbol{E}}_f.$$

For $s \le S$, the source predictor is the same on both target folds, so

$$\begin{aligned}\widehat{g}_s &= \frac{1}{n_T}(\widetilde{\boldsymbol{\Delta}}_s + \boldsymbol{\xi}_s)^\top \boldsymbol{X}_T\boldsymbol{\varepsilon}_T \\ &= \frac{1}{n_T}\widetilde{\boldsymbol{\Delta}}_s^\top \boldsymbol{X}_T\boldsymbol{\varepsilon}_T + (\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^\circ)^\top \frac{1}{n_T}\boldsymbol{U}_s^\top \boldsymbol{X}_T\boldsymbol{\varepsilon}_T \xrightarrow{a.s.} 0.\end{aligned}$$

Here Lemma 2 applies with $A = \widetilde{\boldsymbol{\Delta}}_s$ and $A = \boldsymbol{U}_s$, and $\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^\circ \xrightarrow{a.s.} 0$. For the target coordinate, the same argument gives

$$\begin{aligned}\widehat{g}_{S+1} &= \frac{1}{n_T}\sum_{f=1}^{2}(\boldsymbol{b}_T + \boldsymbol{\xi}_{T,3-f})^\top \boldsymbol{X}_{T,f}\boldsymbol{\varepsilon}_{T,f} \\ &= \sum_{f=1}^{2}\frac{n_{T,f}}{n_T}\left[\frac{1}{n_{T,f}}\boldsymbol{b}_T^\top \boldsymbol{X}_{T,f}\boldsymbol{\varepsilon}_{T,f} + (\widehat{\boldsymbol{\beta}}_{T,3-f} - \boldsymbol{\beta}_T^\circ)^\top \frac{1}{n_{T,f}}\boldsymbol{U}_T^\top \boldsymbol{X}_{T,f}\boldsymbol{\varepsilon}_{T,f}\right] \xrightarrow{a.s.} 0.\end{aligned}$$

Thus $\widehat{\boldsymbol{g}} \xrightarrow{a.s.} 0$.

We next consider $\widehat{\boldsymbol{H}}$. For $s, s' \le S$,

$$\begin{aligned}\widehat{H}_{ss'} &= (\widetilde{\boldsymbol{\Delta}}_s + \boldsymbol{\xi}_s)^\top \widehat{\boldsymbol{\Sigma}}_T(\widetilde{\boldsymbol{\Delta}}_{s'} + \boldsymbol{\xi}_{s'}) \\ &= \widetilde{\boldsymbol{\Delta}}_s^\top \widehat{\boldsymbol{\Sigma}}_T\widetilde{\boldsymbol{\Delta}}_{s'} + \widetilde{\boldsymbol{\Delta}}_s^\top \widehat{\boldsymbol{\Sigma}}_T\boldsymbol{\xi}_{s'} + \boldsymbol{\xi}_s^\top \widehat{\boldsymbol{\Sigma}}_T\widetilde{\boldsymbol{\Delta}}_{s'} + \boldsymbol{\xi}_s^\top \widehat{\boldsymbol{\Sigma}}_T\boldsymbol{\xi}_{s'},\end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}_T = \boldsymbol{X}_T\boldsymbol{X}_T^\top/n_T$. Applying Lemma 1 to coordinate vectors gives $\widehat{\boldsymbol{\Sigma}}_T \xrightarrow{a.s.} \boldsymbol{\Sigma}_T$ in operator norm because D is fixed. Each term involving an estimation error vanishes. For example,

$$|\widetilde{\boldsymbol{\Delta}}_s^\top \widehat{\boldsymbol{\Sigma}}_T\boldsymbol{\xi}_{s'}| \le \|\widetilde{\boldsymbol{\Delta}}_s\|_2\|\widehat{\boldsymbol{\Sigma}}_T\|_{\mathrm{op}}\|\boldsymbol{\xi}_{s'}\|_2 \xrightarrow{a.s.} 0, \qquad |\boldsymbol{\xi}_s^\top \widehat{\boldsymbol{\Sigma}}_T\boldsymbol{\xi}_{s'}| \le \|\boldsymbol{\xi}_s\|_2\|\widehat{\boldsymbol{\Sigma}}_T\|_{\mathrm{op}}\|\boldsymbol{\xi}_{s'}\|_2 \xrightarrow{a.s.} 0.$$

Consequently,

$$\widehat{H}_{ss'} \xrightarrow{a.s.} \widetilde{\boldsymbol{\Delta}}_s^\top \boldsymbol{\Sigma}_T\widetilde{\boldsymbol{\Delta}}_{s'}.$$

For $s \le S$,

$$\widehat{H}_{s,S+1} = \sum_{f=1}^{2}\frac{n_{T,f}}{n_T}(\widetilde{\boldsymbol{\Delta}}_s + \boldsymbol{\xi}_s)^\top \widehat{\boldsymbol{\Sigma}}_{T,f}(\boldsymbol{b}_T + \boldsymbol{\xi}_{T,3-f}),$$

where $\widehat{\boldsymbol{\Sigma}}_{T,f} = \boldsymbol{X}_{T,f}\boldsymbol{X}_{T,f}^\top/n_{T,f}$. For each fold, expanding this product gives

$$\begin{aligned}&(\widetilde{\boldsymbol{\Delta}}_s + \boldsymbol{\xi}_s)^\top \widehat{\boldsymbol{\Sigma}}_{T,f}(\boldsymbol{b}_T + \boldsymbol{\xi}_{T,3-f}) \\ &\quad= \widetilde{\boldsymbol{\Delta}}_s^\top \widehat{\boldsymbol{\Sigma}}_{T,f}\boldsymbol{b}_T + \widetilde{\boldsymbol{\Delta}}_s^\top \widehat{\boldsymbol{\Sigma}}_{T,f}\boldsymbol{\xi}_{T,3-f} + \boldsymbol{\xi}_s^\top \widehat{\boldsymbol{\Sigma}}_{T,f}\boldsymbol{b}_T + \boldsymbol{\xi}_s^\top \widehat{\boldsymbol{\Sigma}}_{T,f}\boldsymbol{\xi}_{T,3-f} \\ &\quad\xrightarrow{a.s.} \widetilde{\boldsymbol{\Delta}}_s^\top \boldsymbol{\Sigma}_T\boldsymbol{b}_T.\end{aligned}$$

The error terms vanish by the same norm bounds, since $\widehat{\boldsymbol{\Sigma}}_{T,f} \xrightarrow{a.s.} \boldsymbol{\Sigma}_T$ and $\boldsymbol{\xi}_{T,3-f} \xrightarrow{a.s.} 0$. Using $\sum_f n_{T,f}/n_T = 1$ gives

$$\widehat{H}_{s,S+1} \xrightarrow{a.s.} \widetilde{\boldsymbol{\Delta}}_s^\top \boldsymbol{\Sigma}_T\boldsymbol{b}_T.$$

Finally,

$$\begin{aligned}\widehat{H}_{S+1,S+1} &= \sum_{f=1}^{2} \frac{n_{T,f}}{n_T} (\boldsymbol{b}_T + \boldsymbol{\xi}_{T,3-f})^\top \widehat{\boldsymbol{\Sigma}}_{T,f} (\boldsymbol{b}_T + \boldsymbol{\xi}_{T,3-f}) \\ &= \sum_{f=1}^{2} \frac{n_{T,f}}{n_T} \left[\boldsymbol{b}_T^\top \widehat{\boldsymbol{\Sigma}}_{T,f} \boldsymbol{b}_T + 2\boldsymbol{b}_T^\top \widehat{\boldsymbol{\Sigma}}_{T,f} \boldsymbol{\xi}_{T,3-f} + \boldsymbol{\xi}_{T,3-f}^\top \widehat{\boldsymbol{\Sigma}}_{T,f} \boldsymbol{\xi}_{T,3-f} \right] \\ &\xrightarrow{a.s.} \boldsymbol{b}_T^\top \boldsymbol{\Sigma}_T \boldsymbol{b}_T.\end{aligned}$$

These entrywise limits give

$$\widehat{\boldsymbol{H}} \xrightarrow{a.s.} \begin{bmatrix} \widetilde{\boldsymbol{\Delta}}_1 & \cdots & \widetilde{\boldsymbol{\Delta}}_S & \boldsymbol{b}_T \end{bmatrix}^\top \boldsymbol{\Sigma}_T \begin{bmatrix} \widetilde{\boldsymbol{\Delta}}_1 & \cdots & \widetilde{\boldsymbol{\Delta}}_S & \boldsymbol{b}_T \end{bmatrix} = \boldsymbol{Q}^\top \boldsymbol{\Sigma}_T \boldsymbol{Q} = \overline{\boldsymbol{H}}.$$

The limiting weights minimize $\boldsymbol{w}^\top \overline{\boldsymbol{H}} \boldsymbol{w}$ subject to $\boldsymbol{w} \geq 0$ and $\mathbf{1}^\top \boldsymbol{w} = 1$. The feasible set is compact and the objective is strictly convex because $\overline{\boldsymbol{H}} \succ 0$, so its minimizer $\boldsymbol{w}^\star$ exists and is unique. Every feasible vector satisfies $\|\boldsymbol{w}\|_2 \leq 1$, giving the uniform bound

$$\sup_{\substack{\boldsymbol{w} \geq 0 \\ \mathbf{1}^\top \boldsymbol{w} = 1}} \left| L(\boldsymbol{w}) - \frac{1}{n_T} \|\boldsymbol{\varepsilon}_T\|_2^2 - \boldsymbol{w}^\top \overline{\boldsymbol{H}} \boldsymbol{w} \right| \leq 2\|\widehat{\boldsymbol{g}}\|_2 + \|\widehat{\boldsymbol{H}} - \overline{\boldsymbol{H}}\|_{\mathrm{op}} \xrightarrow{a.s.} 0.$$

To show that the empirical minimizer converges, take any feasible $\boldsymbol{w}$. The segment from $\boldsymbol{w}^\star$ to $\boldsymbol{w}$ is feasible, and optimality at $\boldsymbol{w}^\star$ makes the directional derivative nonnegative. Thus $(\boldsymbol{w} - \boldsymbol{w}^\star)^\top \overline{\boldsymbol{H}} \boldsymbol{w}^\star \geq 0$, and

$$\begin{aligned}\boldsymbol{w}^\top \overline{\boldsymbol{H}} \boldsymbol{w} - (\boldsymbol{w}^\star)^\top \overline{\boldsymbol{H}} \boldsymbol{w}^\star &= 2(\boldsymbol{w} - \boldsymbol{w}^\star)^\top \overline{\boldsymbol{H}} \boldsymbol{w}^\star + (\boldsymbol{w} - \boldsymbol{w}^\star)^\top \overline{\boldsymbol{H}} (\boldsymbol{w} - \boldsymbol{w}^\star) \\ &\geq \lambda_{\min}(\overline{\boldsymbol{H}}) \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2.\end{aligned}$$

Since $L(\widehat{\boldsymbol{w}}) \leq L(\boldsymbol{w}^\star)$, applying the uniform bound to both weights gives

$$\begin{aligned}\lambda_{\min}(\overline{\boldsymbol{H}}) \|\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\|_2^2 &\leq \widehat{\boldsymbol{w}}^\top \overline{\boldsymbol{H}} \widehat{\boldsymbol{w}} - (\boldsymbol{w}^\star)^\top \overline{\boldsymbol{H}} \boldsymbol{w}^\star \\ &\leq 4\|\widehat{\boldsymbol{g}}\|_2 + 2\|\widehat{\boldsymbol{H}} - \overline{\boldsymbol{H}}\|_{\mathrm{op}} \xrightarrow{a.s.} 0.\end{aligned}$$

Because $\lambda_{\min}(\overline{\boldsymbol{H}}) > 0$, this proves

$$\widehat{\boldsymbol{w}} \xrightarrow{a.s.} \boldsymbol{w}^\star = \arg \min_{\substack{\boldsymbol{w} \geq 0 \\ \mathbf{1}^\top \boldsymbol{w} = 1}} \boldsymbol{w}^\top \overline{\boldsymbol{H}} \boldsymbol{w}.$$

Returning to the risk expansion, every weight lies in $[0, 1]$. Terms involving estimation errors therefore vanish. For example,

$$|\widehat{w}_s \widehat{w}_{s'} \widetilde{\boldsymbol{\Delta}}_s^\top \boldsymbol{\Sigma}_T \boldsymbol{\xi}_{s'}| \leq \|\widetilde{\boldsymbol{\Delta}}_s\|_2 \|\boldsymbol{\Sigma}_T\|_{\mathrm{op}} \|\boldsymbol{\xi}_{s'}\|_2 \xrightarrow{a.s.} 0,$$

and terms involving two errors are bounded by $\|\boldsymbol{\Sigma}_T\|_{\mathrm{op}}$ times the product of their norms. Since there are finitely many terms, substituting the weight limits yields

$$\begin{aligned}\mathcal{R}_T(\widehat{\boldsymbol{\theta}}_{\mathrm{AdaST}}) &\xrightarrow{a.s.} \sigma_T^2 + \sum_{s=1}^{S} \sum_{s'=1}^{S} w_s^\star w_{s'}^\star \widetilde{\boldsymbol{\Delta}}_s^\top \boldsymbol{\Sigma}_T \widetilde{\boldsymbol{\Delta}}_{s'} \\ &\quad + 2 \sum_{s=1}^{S} w_s^\star w_{S+1}^\star \widetilde{\boldsymbol{\Delta}}_s^\top \boldsymbol{\Sigma}_T \boldsymbol{b}_T + (w_{S+1}^\star)^2 \boldsymbol{b}_T^\top \boldsymbol{\Sigma}_T \boldsymbol{b}_T \\ &= \sigma_T^2 + (\boldsymbol{w}^\star)^\top \overline{\boldsymbol{H}} \boldsymbol{w}^\star.\end{aligned}$$

The target-only estimator has limiting risk $\sigma_T^2 + \boldsymbol{b}_T^\top \boldsymbol{\Sigma}_T \boldsymbol{b}_T = \sigma_T^2 + \overline{H}_{S+1,S+1}$. Since $\mathbf{e}_{S+1}$ is feasible, optimality gives

$$(\boldsymbol{w}^\star)^\top \overline{\boldsymbol{H}} \boldsymbol{w}^\star \leq \mathbf{e}_{S+1}^\top \overline{\boldsymbol{H}} \mathbf{e}_{S+1} = \overline{H}_{S+1,S+1}.$$

Equality holds exactly when $\boldsymbol{w}^\star = \mathbf{e}_{S+1}$, since the limiting minimizer is unique. This completes the proof.

E FINITE-SAMPLE DECAY OF THE TARGET WEIGHT

We quantify the **finite-sample** error under the Gaussian-design and fixed-subspace assumptions of Theorem 1. Write $\widehat{w}_T = \widehat{w}_{S+1}$, $\lambda = \lambda_{\min}(\overline{\boldsymbol{H}}) > 0$, and

$$r_n^2 = \sum_{s=1}^{S} \frac{1}{n_s - k - 1} + \sum_{f=1}^{2} \frac{1}{n_{T,f} - k - 1}.$$

Assume all source and target-fold sample sizes exceed $k+1$. Below, C denotes a constant depending only on the fixed population parameters and dimensions.

For source domain s, let $\boldsymbol{\xi}_s = \widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s^\circ$. The projected residual has variance $\nu_s^2 = \sigma_s^2 + \boldsymbol{b}_s^\top \boldsymbol{\Sigma}_s \boldsymbol{b}_s$ and is independent of the projected covariates under Gaussian design. The Gaussian inverse-Gram moment gives

$$\mathbb{E}\|\boldsymbol{\xi}_s\|_2^2 = \frac{\nu_s^2}{n_s - k - 1} \operatorname{tr}\left[(\boldsymbol{U}_s^\top \boldsymbol{\Sigma}_s \boldsymbol{U}_s)^{-1}\right]. \tag{8}$$

The same identity applies to target-fold estimation errors, with $s = T$ and sample size $n_{T,f}$.

Next, use the out-of-fold matrices $\widehat{\boldsymbol{H}}$, $\widehat{\boldsymbol{g}}$, and $\widehat{\boldsymbol{E}}_f$ defined in Appendix D. Set

$$p_f = n_{T,f}/n_T, \qquad \boldsymbol{\Xi}_f = \widehat{\boldsymbol{E}}_f - \boldsymbol{Q}, \qquad \widehat{\boldsymbol{\Sigma}}_{T,f} = \frac{\boldsymbol{X}_{T,f}\boldsymbol{X}_{T,f}^\top}{n_{T,f}}.$$

Since $\boldsymbol{\Xi}_f$ contains source fitting errors and the target fitting error from fold $3-f$, it is independent of the observations in fold f. Expanding $\mathbb{E}\|\boldsymbol{\Xi}_f\|_{\mathrm{F}}^2 = \sum_{s=1}^{S} \mathbb{E}\|\boldsymbol{\xi}_s\|_2^2 + \mathbb{E}\|\boldsymbol{\xi}_{T,3-f}\|_2^2$ and with Equation 8 gives $\mathbb{E}\|\boldsymbol{\Xi}_f\|_{\mathrm{F}}^2 \le C r_n^2$, where $C = \max_s \nu_s^2 \operatorname{tr}\left[(\boldsymbol{U}_s^\top \boldsymbol{\Sigma}_s \boldsymbol{U}_s)^{-1}\right]$. Expanding the quadratic gives

$$\begin{aligned}\widehat{\boldsymbol{H}} - \overline{\boldsymbol{H}} = \boldsymbol{Q}^\top(\widehat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T)\boldsymbol{Q} \\ + \sum_{f=1}^{2} p_f \left[\boldsymbol{Q}^\top \widehat{\boldsymbol{\Sigma}}_{T,f} \boldsymbol{\Xi}_f + \boldsymbol{\Xi}_f^\top \widehat{\boldsymbol{\Sigma}}_{T,f} \boldsymbol{Q} + \boldsymbol{\Xi}_f^\top \widehat{\boldsymbol{\Sigma}}_{T,f} \boldsymbol{\Xi}_f\right].\end{aligned}$$

Since the sample covariance $\widehat{\boldsymbol{\Sigma}}_T$ concentrates around its population counterpart $\boldsymbol{\Sigma}_T$, Gaussian covariance moment bounds yield $\mathbb{E}\left\|\boldsymbol{Q}^\top(\widehat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T)\boldsymbol{Q}\right\| \le C n_T^{-1/2}$. Since $\mathbb{E}\|\widehat{\boldsymbol{\Sigma}}_{T,f}\|_{\mathrm{op}} \le \operatorname{tr}(\boldsymbol{\Sigma}_T)$, then $\mathbb{E}\|\boldsymbol{Q}^\top \widehat{\boldsymbol{\Sigma}}_{T,f} \boldsymbol{\Xi}_f\|_{\mathrm{op}} \le \mathbb{E}\|\boldsymbol{Q}\|_{\mathrm{op}} \operatorname{tr}(\boldsymbol{\Sigma}_T) \sqrt{\mathbb{E}\|\boldsymbol{\Xi}_f\|_{\mathrm{F}}^2} = \mathcal{O}(r_n)$. Similarly, the quadratic term is bounded by $\mathcal{O}(r_n^2)$.

For the variance term,

$$\widehat{\boldsymbol{g}} = \sum_{f=1}^{2} p_f (\boldsymbol{Q} + \boldsymbol{\Xi}_f)^\top \frac{\boldsymbol{X}_{T,f}\boldsymbol{\varepsilon}_{T,f}}{n_{T,f}}, \qquad \mathbb{E}\|\frac{\boldsymbol{X}_{T,f}\boldsymbol{\varepsilon}_{T,f}}{n_{T,f}}\|_2^2 = \frac{\sigma_T^2 \operatorname{tr}(\boldsymbol{\Sigma}_T)}{n_{T,f}}.$$

Similar calculations as for $\widehat{\boldsymbol{H}} - \overline{\boldsymbol{H}}$ give $\mathbb{E}\|\widehat{\boldsymbol{g}}\|_2 \le C(1 + r_n) \sum_f p_f / \sqrt{n_{T,f}}$. Together,

$$\mathbb{E}\left[\|\widehat{\boldsymbol{H}} - \overline{\boldsymbol{H}}\|_{\mathrm{op}} + \|\widehat{\boldsymbol{g}}\|_2\right] \le C(r_n + r_n^2). \tag{9}$$

Finally, let

$$\boldsymbol{d} = \widehat{\boldsymbol{w}} - \boldsymbol{w}^\star.$$

Since $\widehat{\boldsymbol{w}}$ minimizes the empirical quadratic objective over the simplex and $\boldsymbol{w}^\star$ is feasible, the first-order optimality condition gives

$$(\boldsymbol{w}^\star - \widehat{\boldsymbol{w}})^\top \left(\widehat{\boldsymbol{H}}\widehat{\boldsymbol{w}} - \widehat{\boldsymbol{g}}\right) \ge 0,$$

or equivalently,

$$\boldsymbol{d}^\top \left(\widehat{\boldsymbol{H}}\widehat{\boldsymbol{w}} - \widehat{\boldsymbol{g}}\right) \le 0.$$

Similarly, since $\boldsymbol{w}^\star$ minimizes the population quadratic objective over the simplex and $\widehat{\boldsymbol{w}}$ is feasible,

$$(\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star)^\top \overline{\boldsymbol{H}} \boldsymbol{w}^\star \geq 0,$$

so that

$$\boldsymbol{d}^\top \overline{\boldsymbol{H}} \boldsymbol{w}^\star \geq 0.$$

Since $\lambda = \lambda_{\min}(\overline{\boldsymbol{H}})$, we have

$$\begin{aligned}
\lambda \|\boldsymbol{d}\|_2^2 &\leq \boldsymbol{d}^\top \overline{\boldsymbol{H}} \boldsymbol{d} \leq \boldsymbol{d}^\top \overline{\boldsymbol{H}} \widehat{\boldsymbol{w}} \\
&= \boldsymbol{d}^\top \left[(\overline{\boldsymbol{H}} - \widehat{\boldsymbol{H}}) \widehat{\boldsymbol{w}} + \widehat{\boldsymbol{H}} \widehat{\boldsymbol{w}} \right] \\
&= \boldsymbol{d}^\top \left[(\overline{\boldsymbol{H}} - \widehat{\boldsymbol{H}}) \widehat{\boldsymbol{w}} + \widehat{\boldsymbol{g}} \right] + \boldsymbol{d}^\top \left(\widehat{\boldsymbol{H}} \widehat{\boldsymbol{w}} - \widehat{\boldsymbol{g}} \right) \\
&\leq \boldsymbol{d}^\top \left[(\overline{\boldsymbol{H}} - \widehat{\boldsymbol{H}}) \widehat{\boldsymbol{w}} + \widehat{\boldsymbol{g}} \right] \\
&\leq \left(\|\widehat{\boldsymbol{H}} - \overline{\boldsymbol{H}}\|_{\mathrm{op}} + \|\widehat{\boldsymbol{g}}\|_2 \right) \|\boldsymbol{d}\|_2.
\end{aligned}$$

Therefore $\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\|_2 \leq (\|\widehat{\boldsymbol{H}} - \overline{\boldsymbol{H}}\|_{\mathrm{op}} + \|\widehat{\boldsymbol{g}}\|_2)/\lambda$. Taking expectations gives

$$\mathbb{E}\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\|_2 \leq \frac{C}{\lambda}(r_n + r_n^2). \tag{10}$$

In any setting whose population solution has $w_T^\star = 0$, simplex nonnegativity then gives

$$0 \leq \mathbb{E}\widehat{w}_T \leq \frac{C}{\lambda}(r_n + r_n^2). \tag{11}$$

In the empirically-interested regime where source samples are more abundant than target samples, with $n_s \geq c n_T$ for a fixed $c \geq 1$ and $p_f \to \rho_f \in (0,1)$ for $f = 1, 2$, we have $r_n = O(n_T^{-1/2})$. Consequently, if $w_T^\star = 0$,

$$\mathbb{E}\widehat{w}_T = O(n_T^{-1/2}).$$

This establishes a vanishing upper bound on the expected target weight. More generally, the same argument gives $\mathbb{E}|\widehat{w}_T - w_T^\star| = O(n_T^{-1/2})$, including when $w_T^\star > 0$.