SEMANTIC UNCERTAINTY PROBES ACROSS LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic entropy is a useful measure of language-model uncertainty, but computing it requires generating and comparing multiple answers. Semantic entropy probes instead predict this uncertainty from a single forward pass, but they are normally trained separately for each model using expensive semantic entropy labels. We study whether a probe trained on one model can be reused on another. Our method learns a lightweight linear map between paired source and target hidden states, without using target-side semantic entropy labels, and then applies the fixed source probe in the aligned target space. Across 21 ordered source–target pairs spanning different model families and scales, transferred probes recover much of the performance of probes trained directly on the target model and sometimes match or outperform them. To understand these results, we partition prediction outcomes according to source probe errors, transfer-induced changes, and disagreement between source and target uncertainty labels. On average, transfer-induced changes are more often corrective than harmful, while large source–target disagreement limits performance. These findings support the practical reuse of semantic entropy probes and suggest that language models encode partially shared and linearly recoverable uncertainty signals.

1 INTRODUCTION

Large language models (LLMs) have rapidly become a key component of many modern applications and are increasingly relied upon across domains ranging from search engines and social media to law, scientific research, medicine, and chemistry. In many of these settings, an erroneous output can have serious consequences, and the certainty a model conveys matters as much as its accuracy: a model that is confidently wrong is more dangerous than one that signals when it does not know. Reliable uncertainty quantification is therefore becoming a key requirement for trustworthy deployment. It enables systems that are selective in their predictions and can abstain or defer to a human when confidence is low. Yet quantifying the uncertainty of a generative model is particularly challenging, especially compared with classification, where a fixed set of labels and a single correct output make the task much better defined.

Semantic entropy (SE) has emerged as one of the most widely used approaches to this problem (Kuhn et al., 2023). Rather than measuring uncertainty over token sequences, where different sequences may express the same meaning, SE measures uncertainty over semantic equivalence classes. Given an input, one samples multiple generations from the model, clusters them into groups of semantically equivalent answers using a natural language inference model, and computes the entropy of the resulting distribution over clusters. This separates uncertainty about the output's *form* from uncertainty about its *content*.

However, the use of SE comes at a cost. It requires sampling many generations per input and performing semantic clustering at inference time, which is prohibitive in many real-world applications requiring low latency. Semantic entropy probes (SEPs) (Han et al., 2024b) address this by training a lightweight, typically linear model that maps hidden states from a single model run to a prediction of semantic uncertainty. This provides a practical approximation to SE at a fraction of the inference cost. However, SEPs also have limitations. Obtaining the data needed to train them still requires running the full SE pipeline. Moreover, a probe is tied to the representation space of the model on which it was trained, so every new model normally requires a fresh round of label generation and

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.

probe training. This raises the central question of this work: can semantic entropy probes be reused across models?

We answer this question positively through a detailed experimental study. We show that a probe trained on one model can be transferred to another by learning a lightweight adapter that maps hidden states from the target model into the representation space of the source model. The existing source probe can then be applied to the aligned target representations without being retrained. The adapter is learned from paired hidden states on unlabeled inputs and does not require target-side SE labels. Across models spanning different parameter scales and architectural families, we find that transferred probes recover a large fraction of the performance of probes trained directly on the target model. In several cases, the transferred probe matches or outperforms the native target probe.

Previous work has shown that linear features and monitoring probes can sometimes be transferred between model representation spaces using learned linear mappings (Chen et al., 2025; Lee & Kondor, 2026). Our setting introduces an additional difficulty. SE is model-dependent: the source and target models may assign different uncertainty labels to the same input. Transfer therefore requires more than preserving a fixed task label across two representation spaces. We study this setting across independently trained model families and examine how disagreement between their uncertainty labels affects transfer. Due to space constraints, we defer a broader discussion of related work to Appendix A.

The success of transfer is not explained simply by the two models assigning the same uncertainty to every input. Source and target models can disagree on a substantial fraction of their SE labels, yet transfer can still perform well. To understand this, we separate three effects: errors already made by the source probe, changes to its predictions introduced by transfer, and disagreement between the source and target uncertainty labels. This analysis identifies cases in which the transfer corrects an existing error or adapts a prediction to the target model, as well as cases in which it introduces a new error. On average, transfer-induced changes are more often corrective than harmful, while large source-target disagreement limits performance.

Our main contributions are:

- We introduce a simple framework for transferring a semantic entropy probe between language models. The source probe remains fixed, while a lightweight linear adapter is learned from paired, unlabeled hidden representations.
- We evaluate transfer across 21 ordered source-target pairs, covering different model families and parameter scales, as well as different probe and alignment datasets. The transferred probes generally remain close to native target probes without using target-side semantic entropy labels during transfer.
- We analyze cross-model transfer in terms of source probe errors, transfer-induced prediction changes, and disagreement between source and target uncertainty labels. This analysis helps explain both the success and the limitations of transfer.

2 Preliminaries and Problem Setup

2.1 Semantic entropy (SE)

Consider a language model m and an input x. We sample N responses from the model, $y_m^{(1)}, \ldots, y_m^{(N)} \sim p_m(\cdot \mid x)$. Following Kuhn et al. (2023), the responses are grouped into semantic equivalence classes $C_{m,1}(x), \ldots, C_{m,K_m(x)}(x)$ using bidirectional entailment. The empirical probability of class k is

$$\widehat{\pi}_{m,k}(x) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}\left\{ y_m^{(j)} \in C_{m,k}(x) \right\}. \tag{1}$$

SE is the entropy of the resulting distribution over semantic classes:

$$H_m(x) = -\sum_{k=1}^{K_m(x)} \widehat{\pi}_{m,k}(x) \log \widehat{\pi}_{m,k}(x). \tag{2}$$

We use the normalized SE $u_m(x) = \frac{H_m(x)}{\log N} \in [0, 1]$. Higher values indicate that the model produces several semantically different answers, while lower values indicate that its answers are semantically consistent. For the binary task studied in this paper, we define $z_m(x) = \mathbf{1}\{u_m(x) > \gamma_m\}$, where γ_m is the model-specific SE threshold, selected on the training data following the splitting criterion of Han et al. (2024b) to separate low- and high-entropy scores. Related measures, including Kernel Language Entropy and Semantic Volume, use continuous notions of semantic similarity (Nikitin et al., 2024; Li et al., 2026). Our work focuses on the original SE formulation.

2.2 Semantic entropy probes (SEP)

Computing SE requires several generations and a semantic clustering step. A semantic entropy probe (SEP) instead predicts high semantic uncertainty from an internal model representation (Han et al., 2024b). Let $h_m(x) \in \mathbb{R}^{d_m}$ denote the representation extracted from a selected layer and token position of model m. A linear SEP produces the logit

$$a_m(x) = w_m^\top h_m(x) + c_m. \tag{3}$$

The corresponding probability of high semantic uncertainty is $\widehat{p}_m(x) = \sigma(a_m(x))$, where $\sigma(a) = (1 + \exp(-a))^{-1}$. The parameters w_m and c_m are learned by logistic regression using the model-specific labels $z_m(x)$.

We use $\widehat{p}_m(x)$ as the continuous score for Area Under the Receiver Operating Characteristic (AUROC) evaluation. For the binary error analysis, we use the standard logistic regression decision boundary and define $\widehat{z}_m(x) = \mathbf{1}\{\widehat{p}_m(x) > 0.5\}$, equivalently thresholding the probe logit at zero. This threshold is used uniformly throughout.

Once trained, an SEP estimates semantic uncertainty from a single model run. However, generating its training labels still requires the full SE pipeline. The probe is also tied to the hidden space of the model on which it was trained, so its weights cannot in general be applied directly to the representations of another model.

2.3 Cross-model transfer problem

We consider a source model s and a target model t. The source model has a trained SEP with parameters (w_s, c_s). We are also given an unlabeled set of inputs $\{x_i\}_{i=1}^n$, which are passed through both models to obtain the paired representations $\mathcal{A}_{s,t} = \left\{\left(h_s(x_i), h_t(x_i)\right)\right\}_{i=1}^n$.

Our goal is to use these pairs to transfer the fixed source probe to the target model without using target SE labels. Target labels are used only for final evaluation and for training the fully supervised native target baseline.

This problem has two distinct sources of difficulty. First, the source and target representations may have different dimensions and coordinate systems. Second, semantic uncertainty is model-dependent, so the two models may disagree on the same input: $z_s(x) \neq z_t(x)$.

The next section addresses representation mismatch by learning a map from the target hidden space to the source hidden space. We later study how source-target label disagreement limits the resulting transfer.

3 Transferring Semantic Entropy Probes

Section 2.3 introduced a source model s, a target model t, and a source SEP with parameters (w_s, c_s). We now describe how this fixed source probe can be applied to the target model using the paired representations in $\mathcal{A}_{s,t}$.

3.1 Aligned probe transfer

The source probe cannot generally be applied directly to $h_t(x)$ because source and target representations may differ in dimension or coordinate system. We therefore learn an affine map $R_{t\to s}(h) = Mh + b$, where $M \in \mathbb{R}^{d_s \times d_t}$ and $b \in \mathbb{R}^{d_s}$, from the target hidden space to the

source hidden space. The transferred probe logit is

$$a_{s\to t}(x) = w_s^\top R_{t\to s}\big(h_t(x)\big) + c_s = w_s^\top \big(M h_t(x) + b\big) + c_s. \tag{4}$$

The corresponding probability of high semantic uncertainty is $\widehat{p}_{s\to t}(x) = \sigma\left(a_{s\to t}(x)\right)$. For the binary error analysis, we use the standard logistic regression decision boundary and define $\widehat{z}_{s\to t}(x) = \mathbf{1}\{\widehat{p}_{s\to t}(x) > 0.5\}$.

The source probe parameters (w_s, c_s) remain fixed; transfer requires learning only M and b from paired source and target representations.

3.2 Representation alignment

We learn the representation map using ridge-regularized least squares on paired source and target hidden states:

$$(M_{\mathrm{A}}, b_{\mathrm{A}}) = \arg\min_{M,b} \left\{ \frac{1}{n} \sum_{i=1}^{n} \|M h_t(x_i) + b - h_s(x_i)\|_2^2 + \lambda \|M\|_F^2 \right\}, \tag{5}$$

where λ controls the strength of the ridge regularization. We refer to this method as *Affine Alignment*. It learns an affine transformation that places target representations in a coordinate system in which the fixed source probe remains useful. The map is fitted using only the paired representations in $\mathcal{A}_{s,t}$ and does not require target SE labels. Because the source probe is linear, composing the fitted map with w_s is equivalent to fitting a ridge-regularized linear predictor from the target representations to the source probe logits, assuming the same preprocessing and treatment of the intercept. The representation map formulation makes the direction of transfer explicit and allows the source probe to be applied unchanged after alignment.

3.3 Layer selection

The uncertainty signal need not appear at the same depth in different models. We therefore allow $h_s(x)$ and $h_t(x)$ to be extracted from different layers. The source layer is selected during source probe training by maximizing AUROC on source SE labels. For the target layer, we fit an Affine Alignment map for each candidate target layer and select the layer that minimizes the alignment loss, $\|M h_t + b - h_s\|_2^2 + \lambda \|M\|_F^2$. Thus, target layer selection depends only on paired source and target representations and does not require target SE labels. We also test three alternative target layer selection strategies that differ in their use of target-side supervision. A second label-free strategy uses the final transformer layer, motivated by prior observations that SEP performance generally improves at greater model depth (Han et al., 2024b). A full-budget oracle selects the target layer that maximizes probe AUROC using the full target SE label set. Finally, a low-budget variant performs the same AUROC-based layer search using only a small subsample of target SE labels with cross-validation. Results for these three alternative strategies are reported in Appendix B.3.

4 Understanding Cross-Model Transfer

Our experiments in Section 5 show that transferred probes can approach or exceed native target probe AUROC despite using no target SE labels. To understand when transfer succeeds, we examine how source probe errors, changes introduced by transfer, and source-target label disagreement combine to determine the final target prediction.

4.1 Decomposing transfer error

We use the binary labels and predictions defined in Sections 2 and 3. In particular, $z_s(x)$ and $z_t(x)$ are the source and target SE labels, $\widehat{z}_s(x)$ is the native source probe prediction, and $\widehat{z}_{s\to t}(x)$ is the transferred prediction.

For each input x, define the source probe error, transfer-flip, and source-target disagreement indicators, respectively, as

$$\begin{aligned} E_s(x) &= \mathbf{1}\{\widehat{z}_s(x) \neq z_s(x)\}, \quad F_{s\to t}(x) = \mathbf{1}\{\widehat{z}_{s\to t}(x) \neq \widehat{z}_s(x)\}, \\ D_{s,t}(x) &= \mathbf{1}\{z_s(x) \neq z_t(x)\}. \end{aligned} \tag{6}$$

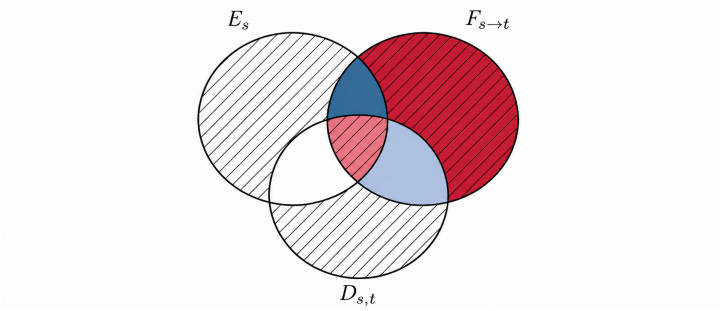

Figure 1: **Decomposition of transfer outcomes.** E_s, $F_{s\to t}$, and $D_{s,t}$ denote source probe error, transfer-induced prediction change, and source–target uncertainty disagreement. Hatched regions are target errors of the transferred probe. Blue regions are corrective transfer flips; red regions are harmful flips.

The final target error of the transferred probe is

$$T_{s\to t}(x) = \mathbf{1}\{\widehat{z}_{s\to t}(x) \neq z_t(x)\} = E_s(x) \oplus F_{s\to t}(x) \oplus D_{s,t}(x), \tag{7}$$

where $\oplus$ denotes exclusive OR. Thus, a transferred prediction is incorrect precisely when an odd number of the three indicators is active:

$$\begin{aligned}\Pr(T_{s\to t}=1) = \; & \Pr(E_s=1, F_{s\to t}=0, D_{s,t}=0) + \Pr(E_s=0, F_{s\to t}=1, D_{s,t}=0) \\ & + \Pr(E_s=0, F_{s\to t}=0, D_{s,t}=1) + \Pr(E_s=1, F_{s\to t}=1, D_{s,t}=1).\end{aligned} \tag{8}$$

These four terms correspond to the regions in which only E_s, only $F_{s\to t}$, only $D_{s,t}$, or all three indicators are active.

4.2 When does transfer help?

The indicator $F_{s\to t}$ identifies exactly the inputs for which transfer changes the binary prediction. If $F_{s\to t}=0$, the source and transferred probes make the same prediction, so transfer does not change target accuracy. If $F_{s\to t}=1$, the change either corrects an existing target error or turns a correct prediction into an error.

Transfer is helpful in two cases. The first occurs when the source probe is wrong about the source model, the source and target labels agree, and transfer corrects the prediction: $E_s=1$, $F_{s\to t}=1$, and $D_{s,t}=0$. The second occurs when the source probe is correct about the source model, the source and target labels disagree, and transfer adapts the prediction to the target: $E_s=0$, $F_{s\to t}=1$, and $D_{s,t}=1$. We therefore define

$$p_c = \Pr(E_s=1, F_{s\to t}=1, D_{s,t}=0) + \Pr(E_s=0, F_{s\to t}=1, D_{s,t}=1). \tag{9}$$

Transfer is harmful in the remaining two cases with $F_{s\to t}=1$. The first occurs when the source prediction is already correct for the target but transfer changes it: $E_s=0$, $F_{s\to t}=1$, and $D_{s,t}=0$. The second occurs when the source probe is wrong about the source model, the source and target labels disagree, and transfer changes a prediction that was already correct for the target: $E_s=1$, $F_{s\to t}=1$, and $D_{s,t}=1$. Thus,

$$p_h = \Pr(E_s=0, F_{s\to t}=1, D_{s,t}=0) + \Pr(E_s=1, F_{s\to t}=1, D_{s,t}=1). \tag{10}$$

The net change in target classification accuracy caused by transfer is

$$\Delta_{\text{acc}} = \Pr(\widehat{z}_{s\to t}=z_t) - \Pr(\widehat{z}_s=z_t) = p_c - p_h. \tag{11}$$

Using the region names reported in our tables in Section 5 and Appendix B, this becomes

$$\Delta_{\text{acc}} = p_{\text{EF only}} + p_{\text{FD only}} - p_{\text{F only}} - p_{\text{EFD}}. \tag{12}$$

A positive value means that transfer corrects more target predictions than it harms. A value close to zero means that the changes introduced by transfer provide little net benefit, while a negative value means that transfer introduces more errors than it corrects.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We train an SEP on source SE labels, fit $R_{t\to s}$ using paired source and target hidden representations, and apply the fixed source probe to $R_{t\to s}(h_t(x))$. Target SE labels are used exclusively for final evaluation and for training the fully supervised native target baseline; they are never used to train the source probe, choose the target representation, or fit the alignment map. Throughout, we follow the protocol of Han et al. (2024b) for generation, SE labeling, and probe training.

Models. We study transfer between pairs of open-weight LLMs. Our main pool consists of five models from five prominent families at the 8B–12B scale: Llama-3.1-8B (Grattafiori et al., 2024), Gemma-4-12B (Team et al., 2026), Phi-4 (Abdin et al., 2024), Qwen3-8B (Yang et al., 2025), and Mistral-Nemo-12B (Mistral AI, 2024). From these we form all 20 ordered cross-family pairs. To examine transfer across scale within a single family, we additionally include Llama-3.2-1B (Meta, 2024) and the pair Llama-3.2-1B $\to$ Llama-3.1-8B, giving 21 source–target pairs in total.

As described in Section 3.3, we select the source layer that maximizes source probe AUROC using source SE labels and the target layer that minimizes the alignment loss. Hidden states are extracted at the second-to-last token (SLT) position following Han et al. (2024b).

Data and SE labels. We evaluate on three open-domain QA benchmarks: TriviaQA (Joshi et al., 2017), SQuAD (Rajpurkar et al., 2018), and Natural Questions (NQ) (Kwiatkowski et al., 2019). For each dataset we draw 2000 inputs. Following Han et al. (2024b), we elicit short-form answers by few-shot prompting each model to answer "as briefly as possible" (prompt templates in Han et al., 2024b, Appendix B.1), sample $N{=}10$ responses at temperature $T{=}1.0$, and obtain SE labels by clustering them with DeBERTa-Large (He et al., 2021) and binarizing the resulting entropy. We reserve 500 examples per dataset as a held-out test set and use the remaining 1500 as a training and alignment pool. After hyperparameter selection, the alignment map is fitted on the full 1500-example pool. To study the effect of source-side supervision, we vary the number of source probe training examples, n, within this pool, which we call the *probe budget*.

Unless stated otherwise, the source probe, alignment data, and evaluation data come from the same dataset. We additionally consider all 3×3 combinations of evaluation and alignment dataset. In these cross-dataset experiments, the source probe is trained and evaluated on one dataset while the alignment map is fitted using paired unlabeled representations from another.

Evaluation. We report transferred-probe AUROC against the target SE labels as a function of probe budget n. We compare the transferred probe with a *native target probe*, which is trained directly on target SE labels. It provides a supervised reference for the performance attainable when target labels are available. We also report the *native source probe* evaluated against source labels. This measures the quality of the original probe before transfer. All probes are logistic regression classifiers (L-BFGS, default ℓ_2 regularization). For the alignment map, we use the Affine Alignment objective of Section 3.2. We sweep the regularization parameter $\lambda \in \{10^1, 10^2, 10^3, 10^4, 10^5\}$ and select $\lambda = 10^4$ based on downstream transfer AUROC. We use this value throughout the main experiments; Table 3 in Appendix B.1 reports the sensitivity of downstream AUROC to λ.

5.2 RESULTS

5.2.1 TRANSFER PERFORMANCE

Figure 2 reports AUROC as a function of the probe budget n for six representative source–target pairs on TriviaQA; results for all 21 pairs and the other evaluation datasets are provided in Appendix B.2.2. Transferred probes remain competitive with native target probes for many model pairs, and in several cases match or exceed them. Effective transfer therefore does not require target SE labels for target layer selection. We use alignment-based target layer selection described in Section 3.3 throughout the main experiments. Appendix B.3 compares this approach with the three alternative strategies described there. The main qualitative transfer patterns persist across all four strategies, although absolute performance varies.

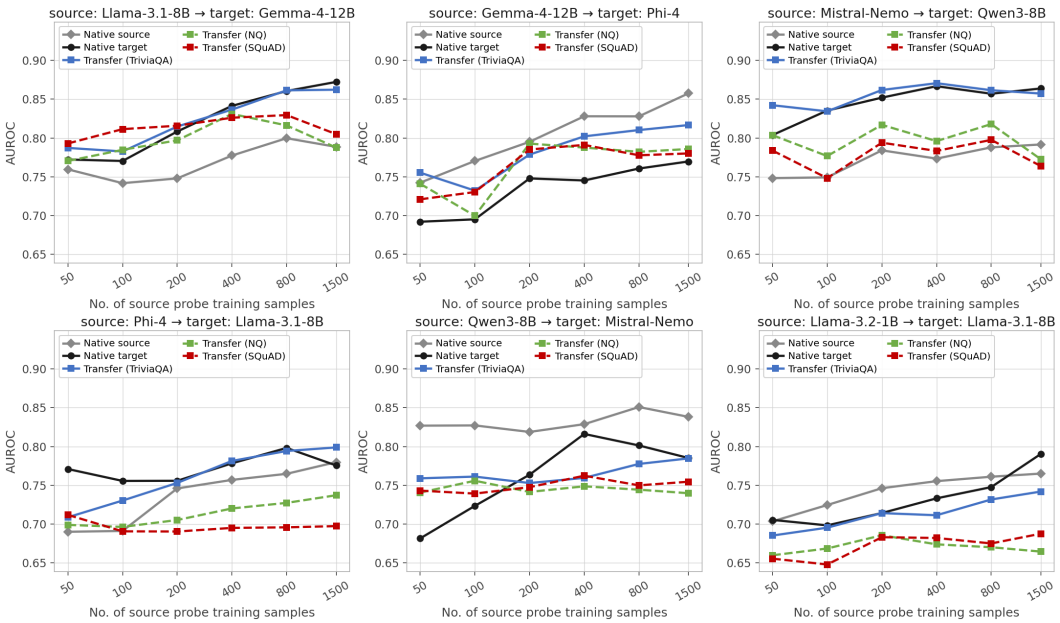

Figure 2: AUROC as a function of the probe budget (n source probe training samples) on TriviaQA for 6 source→target pairs. *Native target* is the target probe trained directly on n target SE labels; *native source* is the source probe trained on n source SE labels and scored against source SE labels. The transferred probe uses Affine Alignment (λ=10^4) from the target layer with the lowest alignment objective to the selected source probe layer. Alignment data are drawn from TriviaQA, NQ, or SQuAD. TriviaQA gives same-dataset alignment, while NQ and SQuAD give cross-dataset alignment.

AUROC generally improves as the source probe budget n increases, with some non-monotonicity at the smallest budgets. Since the alignment map uses the same unlabeled alignment pool independently of n, changes with n primarily reflect the quality of the source probe. Transfer performance also depends substantially on the source–target direction: opposite directions between the same two model families can behave differently, suggesting that successful transfer depends on more than overall similarity between their representation spaces.

Fitting the alignment map on a different dataset consistently degrades performance. In Figure 2, the probe is trained and evaluated on TriviaQA while the alignment map is fitted either on TriviaQA or on NQ and SQuAD. Same-dataset alignment performs best on average, although useful transfer can remain under dataset mismatch. The corresponding SQuAD and NQ results show the same general pattern (Figures 4 and 5 in Appendix B.2.2).

5.2.2 Error decomposition

AUROC measures ranking performance but does not reveal how transfer changes individual binary predictions. We therefore use the decomposition of Section 4 to examine when alignment corrects or introduces errors. Table 1 reports the four regions in which transfer changes the prediction for all 21 source–target pairs at n=1500 on TriviaQA. The two corrective regions, EF_{only} and FD_{only}, account for 11.6% of inputs on average, compared with 5.4% in the pure transfer-harm region F_{only}. Excluding the triple-interaction region gives a mean corrective margin $\Delta_{\text{excl}} = +6.1$ percentage points, which is positive for all 21 pairs.

The EFD region accounts for a further 2.8% of inputs on average and represents an additional harmful transfer case. Including this region, the exact change in target classification accuracy is $\Delta_{\text{acc}} = +3.3$ percentage points on average and is positive for 18 of the 21 pairs. Thus, on TriviaQA, transfer remains net beneficial on average under the complete binary decomposition, although not for every source–target pair.

Table 1: Error decomposition at n=1500 on TriviaQA; alignment dataset: TriviaQA. EF_{only} and FD_{only} are corrective, whereas F_{only} and EFD are harmful. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$, and $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy.

Source → Target	F_{only}	EF_{only}	FD_{only}	EFD	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	9.4	6.4	3.6	2.6	+0.6	-2.0
Gemma-4-12B → Mistral-Nemo	6.6	6.6	5.2	2.6	+5.2	+2.6
Gemma-4-12B → Phi-4	5.8	6.2	6.4	2.6	+6.8	+4.2
Gemma-4-12B → Qwen3-8B	4.2	4.4	7.6	2.8	+7.8	+5.0
Llama-3.1-8B → Gemma-4-12B	6.0	9.2	2.4	3.0	+5.6	+2.6
Llama-3.1-8B → Mistral-Nemo	4.2	7.6	4.0	3.0	+7.4	+4.4
Llama-3.1-8B → Phi-4	4.6	8.8	4.4	2.2	+8.6	+6.4
Llama-3.1-8B → Qwen3-8B	3.4	8.6	7.6	2.4	+12.8	+10.4
Llama-3.2-1B → Llama-3.1-8B	7.8	8.0	2.6	4.8	+2.8	-2.0
Mistral-Nemo → Gemma-4-12B	7.0	9.0	3.4	2.0	+5.4	+3.4
Mistral-Nemo → Llama-3.1-8B	8.2	7.6	4.0	3.6	+3.4	-0.2
Mistral-Nemo → Phi-4	5.2	8.0	3.6	3.2	+6.4	+3.2
Mistral-Nemo → Qwen3-8B	4.0	8.0	7.0	2.4	+11.0	+8.6
Phi-4 → Gemma-4-12B	5.8	6.6	4.4	2.2	+5.2	+3.0
Phi-4 → Llama-3.1-8B	6.2	6.4	4.4	1.8	+4.6	+2.8
Phi-4 → Mistral-Nemo	4.8	6.6	4.0	2.6	+5.8	+3.2
Phi-4 → Qwen3-8B	4.4	7.0	4.2	2.4	+6.8	+4.4
Qwen3-8B → Gemma-4-12B	4.2	6.0	4.0	3.0	+5.8	+2.8
Qwen3-8B → Llama-3.1-8B	4.8	5.8	3.2	3.6	+4.2	+0.6
Qwen3-8B → Mistral-Nemo	2.8	5.4	5.4	3.6	+8.0	+4.4
Qwen3-8B → Phi-4	4.6	6.6	2.2	3.2	+4.2	+1.0
Mean	**5.4**	**7.1**	**4.5**	**2.8**	**+6.1**	**+3.3**

Full error decompositions across all source–target pairs, evaluation datasets, and alignment datasets are provided in Appendix B.2.1. Transfer becomes more difficult when source and target SE labels disagree strongly. The cross-scale Llama-3.2-1B→Llama-3.1-8B pair has particularly high label disagreement and remains a difficult TriviaQA transfer, with a transfer error of 50.6%.

Cross-dataset alignment increases harmful prediction changes substantially. On TriviaQA, mean transfer error rises from 23.9% under TriviaQA alignment to 36.1% under NQ alignment and 43.1% under SQuAD alignment. Correspondingly, Δ_{excl} changes from $+6.1$ to -4.6 and -11.1, while the exact changes are $\Delta_{\text{acc}} = +3.3$, -8.9, and -15.9 percentage points, respectively.

The same-dataset alignment experiments on SQuAD and NQ further show that the effect varies across evaluation datasets. The mean Δ_{excl} remains positive on both datasets ($+3.9$ points on SQuAD and $+8.4$ points on NQ). After accounting for the EFD interaction, the exact mean changes are $+0.1$ points on SQuAD and $+3.6$ points on NQ. Thus, transfer-induced changes contain a consistent corrective component under same-dataset alignment. When interactions with source–target label disagreement are included, this advantage remains substantial on NQ ($+3.6$ points) but is nearly neutral on SQuAD ($+0.1$ points).

The effect of dataset mismatch is substantially stronger. On SQuAD, Δ_{acc} falls from $+0.1$ under SQuAD alignment to -1.8 under NQ alignment and -14.1 under TriviaQA alignment. On NQ, it falls from $+3.6$ under NQ alignment to -0.6 under SQuAD alignment and -7.8 under TriviaQA alignment. These results indicate that the learned correspondence is substantially more reliable when the alignment data match the deployment dataset.

5.2.3 Computational cost

Table 2 reports the wall-clock cost of training a target-side probe from scratch versus transferring an existing source probe. Training a native target probe requires target SE labels, and obtaining these labels through answer sampling and semantic clustering accounts for over 99% of the cost. Transferring an existing source probe requires only target hidden-state extraction and fitting the lightweight alignment map, with alignment fitting accounting for less than 1% of the transfer cost. The resulting end-to-end saving is approximately $12\times$: the mean speedups are $11.86\times$ on NQ, $11.83\times$ on SQuAD, and $11.70\times$ on TriviaQA. This comparison assumes that the source probe already exists,

Table 2: Wall-clock cost of training a target-side probe from scratch versus transferring an existing source probe, broken down by stage (seconds, single GPU). Target-trained stages depend only on the target model, whereas alignment fitting is averaged over source models paired with that target. Probe training and alignment fitting are single fits at n=1500. Transfer replaces answer sampling and semantic clustering, yielding an approximately $12\times$ end-to-end saving.

	Target-trained				Transfer (ours)			
Target model	Answer sampling	Semantic clustering	Probe training	Total	Hidden-state extraction	Alignment fitting	Total	Speedup
(a) Natural Questions								
llama-3.1-8b	3735	570	1.85	4307	340	3.88	343	12.54×
qwen3-8b	3900	285	0.80	4186	355	4.04	359	11.67×
phi-4	5565	495	1.45	6061	506	3.08	509	11.91×
gemma-4-12b	7575	420	1.20	7996	689	3.38	692	11.55×
mistral-nemo	5580	480	2.15	6062	507	3.04	510	11.88×
Mean	5271	450	1.49	5722	479	3.49	483	**11.86**×
(b) SQuAD								
llama-3.1-8b	4680	645	0.67	5326	425	3.76	429	12.41×
qwen3-8b	4830	390	0.76	5221	439	3.92	443	11.78×
phi-4	6705	525	1.37	7231	610	3.02	613	11.81×
gemma-4-12b	9120	495	0.74	9616	829	3.31	832	11.55×
mistral-nemo	6150	525	0.84	6676	559	3.02	562	11.88×
Mean	6297	516	0.88	6814	572	3.41	576	**11.83**×
(c) TriviaQA								
llama-3.1-8b	3105	315	1.84	3422	282	3.87	286	11.96×
qwen3-8b	3540	270	2.32	3812	322	4.00	326	11.70×
phi-4	4995	345	3.17	5343	454	3.04	457	11.69×
gemma-4-12b	6465	330	2.28	6797	588	3.31	591	11.50×
mistral-nemo	4770	375	3.23	5148	434	3.06	437	11.79×
Mean	4575	327	2.57	4905	416	3.46	419	**11.70**×

corresponding to the intended setting in which its label-acquisition cost is amortized across multiple unlabeled target models.

6 Discussion and Conclusion

We studied whether semantic entropy probes can be reused across language models without generating new target-side SE labels. A simple affine alignment of hidden representations is often sufficient to transfer a fixed source probe across model families and scales. Across 21 source–target pairs, transferred probes recover much of the performance of native target probes and sometimes match or exceed them. The error decomposition further shows that transfer-induced changes are, on average, more often corrective than harmful. Together, these results suggest that uncertainty-related information learned by SEPs is at least partially recoverable across model representation spaces.

These findings should not be interpreted as evidence that different language models share a single common uncertainty representation. Successful transfer may instead reflect a lower-dimensional uncertainty-related signal that can be recovered through linear alignment. Moreover, transfer becomes less reliable when source and target uncertainty labels disagree strongly, and alignment learned on a different dataset degrades substantially. Our experiments are also limited to open-weight models, short-form question answering, and affine representation maps.

Several directions are therefore natural for future work. More data-efficient or nonlinear alignment methods may improve transfer when model representations differ substantially, while alignment objectives that explicitly preserve probe-relevant directions may avoid fitting unnecessary parts of the representation space. It would also be useful to study whether uncertainty probes can transfer across broader model scales, tasks, and forms of uncertainty, and whether a single source probe or shared alignment can support many target models. More broadly, understanding which uncertainty-related features are genuinely shared across models may help make uncertainty monitoring substantially cheaper and more reusable.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation of methods and the interpretation of results. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for or assist in writing mathematical proofs, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, assist with translation, clean or reformat datasets, or support qualitative or thematic data analysis; these tasks were either performed without generative AI assistance or were not applicable to this work.

Additionally, we used generative AI tools to create or modify scientific figures or images, draft parts of the research paper, summarize or analyze existing literature, improve the readability of the manuscript, and identify relevant literature. We reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. Interpretations of results were independently assessed against the underlying data and analyses, and AI-assisted literature summaries and identified references were checked against the original sources. All AI-assisted text and figures were reviewed and revised for accuracy, clarity, and consistency with the work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide anonymized code for reproducing our experiments as supplementary material. The experimental setup and implementation details are described in Section 5.1, with additional results and analyses provided in Appendix B.

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.

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023.

David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1881–1890, 2018.

Yixin Bu, Guanyun Zou, Renzhi Wang, Runze Xia, Cunjun Wang, Hongliang Dai, Xiaoqing Ma, and Piji Li. Sampling-free uncertainty quantification via hidden state dynamics in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30104–30111, 2026.

Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. Transferring linear features across language models with model stitching. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Yue Ding, Xiaofang Zhu, Tianze Xia, Junfei Wu, Xinlong Chen, Qiang Liu, and Liang Wang. D^2HScore: Reasoning-aware hallucination detection via semantic breadth and depth analysis in llms. *arXiv preprint arXiv:2509.11569*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiatong Han, Jannik Kossen, Muhammed Razzak, and Yarin Gal. Semantic entropy neurons: Encoding semantic uncertainty in the latent space of LLMs. In *Advances in Neural Information Processing Systems (NeurIPS) - Workshop on Foundation Model Interventions (MINT)*, 2024a.

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024b.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Christopher Irwin, Francesco Caso, Riccardo Ali, and Pietro Lio. Entropy-lens: Uncovering decision strategies in LLMs. In *ICLR 2026 Workshop Sci4DL*, 2026.

Mandar Joshi, Eunsol Choi, Daniel S 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 (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Sunwoo Kim, Haneul Yoo, and Alice Oh. On the effect of uncertainty on layer-wise inference dynamics. In *ICML 2025 Workshop on Actionable Interpretability*, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, pp. 3519–3529. PMLR, 2019.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Su Hyeong Lee and Risi Kondor. Deep minds and shallow probes. *arXiv preprint arXiv:2605.11448*, 2026.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Xiaomin Li, Zhou Yu, Ziji Zhang, Yingying Zhuang, Swair Shah, Narayanan Sadagopan, and Anurag Beniwal. Semantic volume: Quantifying and detecting both external and internal uncertainty in LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31751–31759, 2026.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023.

Meta. Llama 3.2 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md, 2024.

Mistral AI. Mistral nemo. https://mistral.ai/news/mistral-nemo/, 2024.

Zakhar Mrykhin and Valentin Malykh. Prompt embedding probes (PEP): Hallucination detection in LLMs from hidden states. *arXiv preprint arXiv:2608.08024*, 2026.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 8901–8929, 2024.

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. Steer LLM latents for hallucination detection. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in Neural Information Processing Systems*, 30, 2017.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, et al. A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in llm outputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35700–35719, 2025.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33:22045–22055, 2020.

Ponhvoan Srey, Xiaobao Wu, Cong-Duy T Nguyen, and Luu Anh Tuan. Learning uncertainty from sequential internal dispersion in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40088–40106, 2026.

Qiyao Sun, Xingming Li, Xixiang He, Ao Cheng, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Efficient hallucination detection: Adaptive bayesian estimation of semantic entropy with guided semantic exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33117–33125, 2026.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, et al. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for LLMs. *arXiv preprint arXiv:2505.20045*, 2025.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F. Wong, and Rui Wang. Latent space chain-of-embedding enables output-free LLM self-evaluation. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775. PMLR, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Related Work

A.1 Semantic uncertainty

Token-level predictive entropy is difficult to interpret for free-form generation because many lexically different responses can express the same answer. Semantic entropy addresses this problem by grouping sampled responses according to their meaning and measuring uncertainty over the resulting semantic classes (Kuhn et al., 2023). The original formulation treats semantic equivalence classes as discrete. Subsequent work develops finer-grained measures of semantic variation. Kernel Language Entropy incorporates pairwise semantic similarities between generations (Nikitin et al., 2024), while Semantic Volume measures the dispersion of semantic representations (Li et al., 2026). These approaches enrich the representation of semantic uncertainty but still rely on multiple generations at inference time. Our work uses the standard semantic-entropy construction and focuses on reducing the cost of deploying a predictor of this quantity on new models.

A.2 Efficient uncertainty estimation

Methods for estimating LLM uncertainty differ in both their inference-time computational requirements and their reliance on labeled training data. Sampling-based methods, including the semantic-uncertainty approaches discussed above, estimate uncertainty from variation across multiple generations. Semantic entropy measures this variation over semantic equivalence classes (Kuhn et al., 2023), while INSIDE measures it in the internal representation space using the covariance of hidden-state embeddings (Chen et al., 2024). Recent work reduces the sampling cost of semantic entropy through adaptive sampling and guided semantic exploration (Sun et al., 2026). Nevertheless, these methods still require repeated generation for each query.

Sampling-free methods instead derive uncertainty signals from a single generation. Several do so without task-specific supervision. Sampling-Free UQ models the evolution of hidden states during generation (Bu et al., 2026), while D^2HScore characterizes intra-layer dispersion and inter-layer drift (Ding et al., 2025). Other approaches derive uncertainty from attention patterns and token-level confidence (Vazhentsev et al., 2025). Related work also studies how uncertainty is reflected in the evolution of intermediate representations or their induced vocabulary distributions (Wang et al., 2025; Kim et al., 2025; Irwin et al., 2026). More broadly, adjacent work identifies latent directions associated with truthfulness and related properties for hallucination detection or activation steering (Park et al., 2025; Zou et al., 2023; Marks & Tegmark, 2024; Li et al., 2023).

Supervised sampling-free methods learn uncertainty or hallucination predictors from model-internal signals using labeled examples. SIVR learns a sequence classifier over token-wise, cross-layer dispersion features (Srey et al., 2026), while pre-trained UQ heads learn auxiliary predictors from attention and logit features (Shelmanov et al., 2025). Semantic entropy probes (SEPs) instead train linear probes on hidden representations using binarized semantic-entropy labels (Han et al., 2024b). Although an SEP requires only a single generation at inference time, obtaining its training labels still requires the sampling-based SE procedure. Semantic Entropy Neurons further studies the localization of this signal, showing that SE can be predicted from a small subset of latent dimensions (Han et al., 2024a). Prior work trains and evaluates probes within the representation space of the same underlying model. Our work instead asks whether an SEP trained for one model can be reused on another.

Prompt Embedding Probes (PEP) extend hidden-state probing with learned prompt embeddings and evaluate cross-model transfer using a fixed detector model that classifies text generated by another model (Mrykhin & Malykh, 2026). Such methods transfer at the level of textual outputs: the detector processes only the target model's generated text, not its internal representations. Our setting differs in that we transfer a fixed probe trained specifically on model-dependent semantic-entropy labels by explicitly aligning the representation spaces of source and target models, without requiring the source model to process the target model's generated text at deployment.

A.3 Representation alignment and probe transfer

Comparing and aligning neural representations has a long history. Canonical Correlation Analysis and its variants identify correlated linear subspaces (Raghu et al., 2017), while Centered Kernel

Alignment provides a similarity measure that is invariant to several transformations of the representation space (Kornblith et al., 2019). Optimal transport methods can align representations or neurons across networks, including settings in which their internal structures differ (Alvarez-Melis & Jaakkola, 2018; Singh & Jaggi, 2020). Permutation-based alignment has similarly been used to bring independently trained networks into correspondence before model merging (Ainsworth et al., 2023). Adjacent work operates directly in parameter space, including model merging (Yadav et al., 2023; Yu et al., 2024), task-vector arithmetic (Ilharco et al., 2023), and model editing (Meng et al., 2023).

Most closely related to our method, Chen et al. (2025) study the transfer of linear features between language models through affine mappings of their residual-stream representations. They show that learned mappings can transfer probes and steering directions, particularly between models from related families. Lee & Kondor (2026) study the transfer of linear monitoring probes across language models using paired, unlabeled activations and identify conditions under which representation alignment preserves probe-relevant directions. They show that successful probe transfer depends on preserving probe-relevant information, rather than simply achieving low alignment error over the full representation.

These works establish that linear probes can sometimes be transferred between model representation spaces using unlabeled paired activations. Our setting introduces an additional difficulty: the target quantity itself is model-dependent. Semantic-entropy labels are generated from each model's own response distribution, so the source and target labels may disagree even when both are evaluated on the same input. We study how alignment performs under this model-dependent supervision and how label disagreement interacts with the changes introduced by transfer. Our experiments also cover transfers across independently trained architectural families, model scales, and datasets.

B Additional Experimental Results

This appendix provides additional results supporting Section 5. Section B.1 examines sensitivity to alignment regularization. Section B.2 reports the complete error decompositions and transfer curves for our main results across all evaluation–alignment dataset combinations and all 21 source–target pairs. Finally, Section B.3 evaluates the three alternative target layer selection strategies introduced in Section 3.3.

B.1 Regularization Sensitivity

The main experiments use $\lambda = 10^4$. We tune the regularization strength on NQ by sweeping $\lambda \in \{10^1, 10^2, 10^3, 10^4, 10^5\}$ and selecting $\lambda = 10^4$ based on transfer AUROC on the NQ test set. We then fix this value for all other datasets without dataset-specific retuning.

Table 3 reports the resulting NQ sensitivity analysis under same-dataset alignment, averaged over all 21 source–target pairs. Performance is evaluated across the full six-point probe-budget grid, and average rank summarizes performance across pairs and probe budgets, with lower rank indicating better overall performance.

B.2 Full Results with Alignment-Based Target Layer Selection

We report complete results for the target layer selection strategy used in the main experiments, covering all evaluation–alignment dataset combinations and all 21 source–target pairs.

Table 3: Regularization sensitivity analysis for Affine Alignment on NQ, showing AUROC averaged over all 21 source–target pairs under same-dataset alignment. The alignment budget is fixed at 1500. Avg. rank is computed over the full 6-point probe-budget grid and all pairs; lower is better.

		source probe training samples n				
Method	Reg.	50	200	800	1500	Avg. rank
References						
Native target probe (needs n target labels)	–	.671	.714	.753	.765	–
Native source probe	–	.669	.714	.752	.764	–
Affine Alignment	λ=10^1	.596	.646	.668	.684	4.86
	λ=10^2	.611	.671	.693	.711	3.79
	λ=10^3	.632	.713	.738	.758	2.52
	λ=10^4	**.642**	**.745**	**.770**	**.786**	**1.26**
	λ=10^5	.615	.737	.757	.767	2.57

B.2.1 Full error decompositions

Tables 4–12 report the complete decomposition of Section 4.1 for all evaluation–alignment dataset combinations. Each table reports the seven non-empty regions induced by E_s, $F_{s\to t}$, and $D_{s,t}$. We also report the transfer error from Equation 8, which is the classification error of the transferred probe with respect to the target SE labels. In the notation of the tables, the transfer error $= E_{\text{only}} + F_{\text{only}} + D_{\text{only}} + EFD$. All entries are percentages of the 500 held-out test examples. We report $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ as the corrective margin excluding the triple-interaction region, and $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ as the exact change in target classification accuracy.

For same-dataset alignment, the mean Δ_{excl} is positive on all three benchmarks: $+6.1$ points on TriviaQA, $+3.9$ points on SQuAD, and $+8.4$ points on NQ. After including the EFD interaction, the corresponding exact mean changes are $+3.3$, $+0.1$, and $+3.6$ points, respectively. Cross-dataset alignment is substantially less reliable.

Table 4: Full error decomposition at n=1500 on TriviaQA; alignment dataset: TriviaQA. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	4.8	9.4	7.8	6.4	5.2	3.6	2.6	24.6	+0.6	-2.0
Gemma-4-12B → Mistral-Nemo	5.0	6.6	8.4	6.6	4.8	5.2	2.6	22.6	+5.2	+2.6
Gemma-4-12B → Phi-4	4.2	5.8	9.6	6.2	6.0	6.4	2.6	22.2	+6.8	+4.2
Gemma-4-12B → Qwen3-8B	2.4	4.2	10.0	4.4	9.4	7.6	2.8	19.4	+7.8	+5.0
Llama-3.1-8B → Gemma-4-12B	5.6	6.0	6.6	9.2	7.2	2.4	3.0	21.2	+5.6	+2.6
Llama-3.1-8B → Mistral-Nemo	7.6	4.2	8.4	7.6	6.8	4.0	3.0	23.2	+7.4	+4.4
Llama-3.1-8B → Phi-4	7.0	4.6	9.4	8.8	7.0	4.4	2.2	23.2	+8.6	+6.4
Llama-3.1-8B → Qwen3-8B	4.4	3.4	9.0	8.6	9.6	7.6	2.4	19.2	+12.8	+10.4
Llama-3.2-1B → Llama-3.1-8B	8.6	7.8	29.4	8.0	2.6	2.6	4.8	50.6	+2.8	-2.0
Mistral-Nemo → Gemma-4-12B	6.2	7.0	7.6	9.0	8.0	3.4	2.0	22.8	+5.4	+3.4
Mistral-Nemo → Llama-3.1-8B	6.8	8.2	7.4	7.6	7.2	4.0	3.6	26.0	+3.4	-0.2
Mistral-Nemo → Phi-4	5.6	5.2	8.8	8.0	8.4	3.6	3.2	22.8	+6.4	+3.2
Mistral-Nemo → Qwen3-8B	3.0	4.0	10.0	8.0	11.8	7.0	2.4	19.4	+11.0	+8.6
Phi-4 → Gemma-4-12B	6.2	5.8	8.6	6.6	9.4	4.4	2.2	22.8	+5.2	+3.0
Phi-4 → Llama-3.1-8B	9.0	6.2	9.6	6.4	7.2	4.4	1.8	26.6	+4.6	+2.8
Phi-4 → Mistral-Nemo	7.8	4.8	10.0	6.6	7.4	4.0	2.6	25.2	+5.8	+3.2
Phi-4 → Qwen3-8B	3.8	4.4	7.0	7.0	11.2	4.2	2.4	17.6	+6.8	+4.4
Qwen3-8B → Gemma-4-12B	4.0	4.2	10.8	6.0	7.2	4.0	3.0	22.0	+5.8	+2.8
Qwen3-8B → Llama-3.1-8B	4.4	4.8	13.0	5.8	6.4	3.2	3.6	25.8	+4.2	+0.6
Qwen3-8B → Mistral-Nemo	4.8	2.8	12.2	5.4	6.4	5.4	3.6	23.4	+8.0	+4.4
Qwen3-8B → Phi-4	3.0	4.6	10.4	6.6	7.4	2.2	3.2	21.2	+4.2	+1.0
Mean	**5.4**	**5.4**	**10.2**	**7.1**	**7.5**	**4.5**	**2.8**	**23.9**	**+6.1**	**+3.3**

Table 5: Full error decomposition at n=1500 on TriviaQA; alignment dataset: NQ. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	5.6	15.6	6.0	5.6	4.8	5.4	3.0	30.2	-4.6	-7.6
Gemma-4-12B → Mistral-Nemo	5.4	17.4	7.6	6.2	4.2	6.0	3.2	33.6	-5.2	-8.4
Gemma-4-12B → Phi-4	4.2	18.8	9.0	6.2	5.4	7.0	3.2	35.2	-5.6	-8.8
Gemma-4-12B → Qwen3-8B	2.8	18.0	8.0	4.0	7.0	9.6	5.2	34.0	-4.4	-9.6
Llama-3.1-8B → Gemma-4-12B	7.4	25.0	5.8	7.4	5.2	3.2	5.0	43.2	-14.4	-19.4
Llama-3.1-8B → Mistral-Nemo	8.4	30.8	6.8	6.8	4.4	5.6	5.4	51.4	-18.4	-23.8
Llama-3.1-8B → Phi-4	8.8	29.0	6.6	7.0	4.8	7.2	4.4	48.8	-14.8	-19.2
Llama-3.1-8B → Qwen3-8B	6.2	27.8	7.8	6.8	7.0	8.8	5.0	46.8	-12.2	-17.2
Llama-3.2-1B → Llama-3.1-8B	10.8	18.2	31.0	5.8	1.2	1.0	6.2	66.2	-11.4	-17.6
Mistral-Nemo → Gemma-4-12B	6.8	10.8	5.8	8.4	7.2	5.2	2.8	26.2	+2.8	0.0
Mistral-Nemo → Llama-3.1-8B	6.8	11.6	8.8	7.6	6.6	2.6	4.2	31.4	-1.4	-5.6
Mistral-Nemo → Phi-4	5.0	8.2	7.6	8.6	7.8	4.8	3.8	24.6	+5.2	+1.4
Mistral-Nemo → Qwen3-8B	4.4	9.0	9.4	6.6	10.0	7.6	4.2	27.0	+5.2	+1.0
Phi-4 → Gemma-4-12B	5.2	11.4	7.6	7.6	8.2	5.4	3.4	27.6	+1.6	-1.8
Phi-4 → Llama-3.1-8B	7.8	20.2	8.0	7.6	6.0	6.0	3.0	39.0	-6.6	-9.6
Phi-4 → Mistral-Nemo	6.8	18.8	7.6	7.6	4.8	6.4	5.2	38.4	-4.8	-10.0
Phi-4 → Qwen3-8B	4.6	17.2	5.6	6.2	8.2	5.6	5.4	32.8	-5.4	-10.8
Qwen3-8B → Gemma-4-12B	4.8	12.8	9.6	5.2	5.8	5.2	4.4	31.6	-2.4	-6.8
Qwen3-8B → Llama-3.1-8B	5.2	12.4	10.4	5.0	6.0	5.8	4.0	32.0	-1.6	-5.6
Qwen3-8B → Mistral-Nemo	5.6	11.6	10.2	4.6	4.4	7.4	5.6	33.0	+0.4	-5.2
Qwen3-8B → Phi-4	4.0	8.8	7.6	5.6	6.2	5.0	4.4	24.8	+1.8	-2.6
Mean	**6.0**	**16.8**	**8.9**	**6.5**	**6.0**	**5.8**	**4.3**	**36.1**	**-4.6**	**-8.9**

Table 6: Full error decomposition at n=1500 on TriviaQA; alignment dataset: SQuAD. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	6.2	21.6	5.2	5.0	4.4	6.2	3.4	36.4	-10.4	-13.8
Gemma-4-12B → Mistral-Nemo	6.0	26.8	8.2	5.6	3.4	5.4	4.0	45.0	-15.8	-19.8
Gemma-4-12B → Phi-4	5.2	27.2	7.8	5.2	4.4	8.2	4.2	44.4	-13.8	-18.0
Gemma-4-12B → Qwen3-8B	2.8	22.0	7.4	4.0	7.2	10.2	5.0	37.2	-7.8	-12.8
Llama-3.1-8B → Gemma-4-12B	8.0	36.6	4.4	6.8	4.8	4.6	5.4	54.4	-25.2	-30.6
Llama-3.1-8B → Mistral-Nemo	8.6	42.4	5.8	6.6	4.4	6.6	5.4	62.2	-29.2	-34.6
Llama-3.1-8B → Phi-4	9.4	40.8	6.0	6.4	3.8	7.8	5.4	61.6	-26.6	-32.0
Llama-3.1-8B → Qwen3-8B	7.2	34.6	7.2	5.8	6.0	9.4	6.0	55.0	-19.4	-25.4
Llama-3.2-1B → Llama-3.1-8B	10.8	17.2	31.0	5.8	1.2	1.0	6.2	65.2	-10.4	-16.6
Mistral-Nemo → Gemma-4-12B	8.2	22.6	6.4	7.0	6.2	4.6	3.8	41.0	-11.0	-14.8
Mistral-Nemo → Llama-3.1-8B	8.8	32.4	5.8	5.6	5.4	5.6	5.4	52.4	-21.2	-26.6
Mistral-Nemo → Phi-4	6.8	24.2	7.0	6.8	6.6	5.4	5.0	43.0	-12.0	-17.0
Mistral-Nemo → Qwen3-8B	5.6	24.6	7.2	5.4	7.8	9.8	6.4	43.8	-9.4	-15.8
Phi-4 → Gemma-4-12B	5.8	21.8	7.4	7.0	7.4	5.6	4.2	39.2	-9.2	-13.4
Phi-4 → Llama-3.1-8B	8.0	25.0	7.4	7.4	4.8	6.6	4.2	44.6	-11.0	-15.2
Phi-4 → Mistral-Nemo	7.4	17.0	8.6	7.0	4.8	5.4	5.2	38.2	-4.6	-9.8
Phi-4 → Qwen3-8B	4.2	16.8	4.8	6.6	9.0	6.4	4.6	30.4	-3.8	-8.4
Qwen3-8B → Gemma-4-12B	2.8	8.4	8.2	7.2	6.8	6.6	3.4	22.8	+5.4	+2.0
Qwen3-8B → Llama-3.1-8B	4.4	10.6	9.2	5.8	5.6	7.0	4.4	28.6	+2.2	-2.2
Qwen3-8B → Mistral-Nemo	4.6	13.8	10.2	5.6	4.8	7.4	5.2	33.8	-0.8	-6.0
Qwen3-8B → Phi-4	3.6	10.0	8.2	6.0	6.2	4.4	4.4	26.2	+0.4	-4.0
Mean	**6.4**	**23.6**	**8.3**	**6.1**	**5.5**	**6.4**	**4.8**	**43.1**	**-11.1**	**-15.9**

Table 7: Full error decomposition at n=1500 on SQuAD; alignment dataset: SQuAD. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	4.0	10.4	8.0	5.8	7.0	4.2	3.8	26.2	-0.4	-4.2
Gemma-4-12B → Mistral-Nemo	5.2	8.4	6.6	7.2	6.4	5.2	1.8	22.0	+4.0	+2.2
Gemma-4-12B → Phi-4	5.4	7.2	8.4	6.6	6.2	6.2	2.4	23.4	+5.6	+3.2
Gemma-4-12B → Qwen3-8B	6.2	11.4	18.4	5.2	4.6	3.2	4.6	40.6	-3.0	-7.6
Llama-3.1-8B → Gemma-4-12B	5.6	5.2	8.0	6.6	9.2	3.0	2.8	21.6	+4.4	+1.6
Llama-3.1-8B → Mistral-Nemo	6.4	4.8	6.6	7.0	8.6	3.6	2.2	20.0	+5.8	+3.6
Llama-3.1-8B → Phi-4	6.2	6.8	9.2	7.4	8.8	3.6	1.8	24.0	+4.2	+2.4
Llama-3.1-8B → Qwen3-8B	8.4	6.6	17.4	6.0	5.2	3.4	4.6	37.0	+2.8	-1.8
Llama-3.2-1B → Llama-3.1-8B	6.2	4.2	13.6	5.2	4.4	1.4	5.8	29.8	+2.4	-3.4
Mistral-Nemo → Gemma-4-12B	7.2	5.6	8.2	6.8	4.6	4.2	3.0	24.0	+5.4	+2.4
Mistral-Nemo → Llama-3.1-8B	6.8	7.8	8.2	5.8	6.0	3.8	3.0	25.8	+1.8	-1.2
Mistral-Nemo → Phi-4	6.2	6.2	7.8	8.2	4.4	3.8	2.8	23.0	+5.8	+3.0
Mistral-Nemo → Qwen3-8B	5.6	7.4	21.2	5.0	4.6	2.6	6.4	40.6	+0.2	-6.2
Phi-4 → Gemma-4-12B	7.4	6.0	8.2	7.6	8.6	2.4	4.0	25.6	+4.0	0.0
Phi-4 → Llama-3.1-8B	7.4	7.0	8.6	7.8	9.4	2.4	3.0	26.0	+3.2	+0.2
Phi-4 → Mistral-Nemo	7.2	6.6	5.0	8.6	8.4	2.0	3.4	22.2	+4.0	+0.6
Phi-4 → Qwen3-8B	9.6	5.0	19.2	6.6	6.8	2.6	4.6	38.4	+4.2	-0.4
Qwen3-8B → Gemma-4-12B	6.4	6.6	14.0	7.2	7.2	4.4	5.2	32.2	+5.0	-0.2
Qwen3-8B → Llama-3.1-8B	5.6	7.0	13.0	7.0	8.4	4.2	5.0	30.6	+4.2	-0.8
Qwen3-8B → Mistral-Nemo	4.6	6.4	15.4	8.6	8.8	6.6	4.0	30.4	+8.8	+4.8
Qwen3-8B → Phi-4	6.0	6.4	14.8	8.4	7.6	6.8	4.0	31.2	+8.8	+4.8
Mean	**6.4**	**6.8**	**11.4**	**6.9**	**6.9**	**3.8**	**3.7**	**28.3**	**+3.9**	**+0.1**

Table 8: Full error decomposition at n=1500 on SQuAD; alignment dataset: NQ. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	2.8	14.6	6.2	7.0	5.8	6.0	5.0	28.6	-1.6	-6.6
Gemma-4-12B → Mistral-Nemo	4.8	10.6	5.8	7.6	4.8	6.0	3.4	24.6	+3.0	-0.4
Gemma-4-12B → Phi-4	5.0	8.6	6.4	7.0	5.4	8.2	3.2	23.2	+6.6	+3.4
Gemma-4-12B → Qwen3-8B	6.0	8.8	15.2	5.4	4.8	6.4	4.4	34.4	+3.0	-1.4
Llama-3.1-8B → Gemma-4-12B	4.8	8.6	8.4	7.4	8.8	2.6	3.2	25.0	+1.4	-1.8
Llama-3.1-8B → Mistral-Nemo	7.0	8.2	6.0	6.4	8.2	4.2	2.6	23.8	+2.4	-0.2
Llama-3.1-8B → Phi-4	6.4	7.2	8.4	7.2	8.6	4.4	2.0	24.0	+4.4	+2.4
Llama-3.1-8B → Qwen3-8B	10.0	9.0	17.6	4.4	4.0	3.2	5.8	42.4	-1.4	-7.2
Llama-3.2-1B → Llama-3.1-8B	5.8	4.8	13.0	5.6	4.4	2.0	5.8	29.4	+2.8	-3.0
Mistral-Nemo → Gemma-4-12B	6.6	7.8	8.2	7.4	4.2	4.2	3.4	26.0	+3.8	+0.4
Mistral-Nemo → Llama-3.1-8B	6.8	8.4	8.8	5.8	6.0	3.2	3.0	27.0	+0.6	-2.4
Mistral-Nemo → Phi-4	6.0	7.2	6.8	8.4	4.8	4.8	2.4	22.4	+6.0	+3.6
Mistral-Nemo → Qwen3-8B	5.4	7.0	20.2	5.2	4.6	3.6	6.4	39.0	+1.8	-4.6
Phi-4 → Gemma-4-12B	7.2	7.6	7.6	7.8	9.4	3.0	3.2	25.6	+3.2	0.0
Phi-4 → Llama-3.1-8B	6.2	7.8	8.2	9.0	8.8	2.8	3.6	25.8	+4.0	+0.4
Phi-4 → Mistral-Nemo	7.2	8.2	4.8	8.6	9.0	2.2	2.8	23.0	+2.6	-0.2
Phi-4 → Qwen3-8B	8.2	4.6	18.6	8.0	5.8	3.2	5.6	37.0	+6.6	+1.0
Qwen3-8B → Gemma-4-12B	6.6	11.2	14.8	7.0	6.6	3.6	5.8	38.4	-0.6	-6.4
Qwen3-8B → Llama-3.1-8B	7.2	13.8	12.6	5.4	6.0	4.6	7.4	41.0	-3.8	-11.2
Qwen3-8B → Mistral-Nemo	4.6	10.2	16.0	8.6	7.4	6.0	5.4	36.2	+4.4	-1.0
Qwen3-8B → Phi-4	7.2	10.6	16.4	7.2	6.6	5.2	5.0	39.2	+1.8	-3.2
Mean	**6.3**	**8.8**	**11.0**	**7.0**	**6.4**	**4.3**	**4.3**	**30.3**	**+2.4**	**-1.8**

Table 9: Full error decomposition at n=1500 on SQuAD; alignment dataset: TriviaQA. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	5.2	39.6	4.2	4.6	4.8	8.0	6.0	55.0	-27.0	-33.0
Gemma-4-12B → Mistral-Nemo	6.4	34.0	5.8	6.0	4.2	6.0	4.0	50.2	-22.0	-26.0
Gemma-4-12B → Phi-4	7.0	37.2	6.6	5.0	3.6	8.0	5.0	55.8	-24.2	-29.2
Gemma-4-12B → Qwen3-8B	3.4	29.2	2.2	8.0	6.6	19.4	2.6	37.4	-1.8	-4.4
Llama-3.1-8B → Gemma-4-12B	5.8	17.8	7.8	6.4	6.0	3.2	6.0	37.4	-8.2	-14.2
Llama-3.1-8B → Mistral-Nemo	6.4	12.0	5.8	7.0	7.6	4.4	3.2	27.4	-0.6	-3.8
Llama-3.1-8B → Phi-4	6.4	15.0	7.2	7.2	8.0	5.6	2.6	31.2	-2.2	-4.8
Llama-3.1-8B → Qwen3-8B	6.4	12.0	12.4	8.0	5.8	8.4	4.0	34.8	+4.4	+0.4
Llama-3.2-1B → Llama-3.1-8B	5.8	4.2	11.8	5.6	5.6	3.2	4.6	26.4	+4.6	0.0
Mistral-Nemo → Gemma-4-12B	7.0	21.0	7.6	7.0	4.8	4.8	2.8	38.4	-9.2	-12.0
Mistral-Nemo → Llama-3.1-8B	7.2	28.6	4.2	5.4	5.6	7.8	3.4	43.4	-15.4	-18.8
Mistral-Nemo → Phi-4	7.8	30.0	5.4	6.6	4.4	6.2	2.8	46.0	-17.2	-20.0
Mistral-Nemo → Qwen3-8B	4.0	15.2	8.2	6.6	8.6	15.6	2.4	29.8	+7.0	+4.6
Phi-4 → Gemma-4-12B	8.4	21.2	5.2	6.6	6.4	5.4	6.2	41.0	-9.2	-15.4
Phi-4 → Llama-3.1-8B	6.0	21.4	4.6	9.2	6.8	6.4	5.6	37.6	-5.8	-11.4
Phi-4 → Mistral-Nemo	7.2	21.0	3.8	8.6	6.0	3.2	5.8	37.8	-9.2	-15.0
Phi-4 → Qwen3-8B	5.8	7.2	11.2	10.4	8.0	10.6	3.4	27.6	+13.8	+10.4
Qwen3-8B → Gemma-4-12B	9.2	27.0	15.0	4.4	2.8	3.4	9.6	60.8	-19.2	-28.8
Qwen3-8B → Llama-3.1-8B	8.6	23.8	13.8	4.0	3.0	3.4	10.4	56.6	-16.4	-26.8
Qwen3-8B → Mistral-Nemo	7.8	27.0	18.6	5.4	3.4	3.4	9.4	62.8	-18.2	-27.6
Qwen3-8B → Phi-4	9.8	21.6	17.6	4.6	3.8	4.0	7.8	56.8	-13.0	-20.8
Mean	**6.7**	**22.2**	**8.5**	**6.5**	**5.5**	**6.7**	**5.1**	**42.6**	**-9.0**	**-14.1**

Table 10: Full error decomposition at n=1500 on NQ; alignment dataset: NQ. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	6.4	7.4	5.8	10.0	10.0	5.6	4.2	23.8	+8.2	+4.0
Gemma-4-12B → Mistral-Nemo	7.6	7.8	9.2	13.4	6.6	5.2	3.0	27.6	+10.8	+7.8
Gemma-4-12B → Phi-4	8.0	6.6	7.6	12.4	7.4	5.2	2.8	25.0	+11.0	+8.2
Gemma-4-12B → Qwen3-8B	5.6	5.6	15.8	10.0	8.6	6.4	6.4	33.4	+10.8	+4.4
Llama-3.1-8B → Gemma-4-12B	6.6	7.4	9.2	11.0	7.2	3.6	5.6	28.8	+7.2	+1.6
Llama-3.1-8B → Mistral-Nemo	6.6	7.2	9.2	11.8	6.8	4.0	5.2	28.2	+8.6	+3.4
Llama-3.1-8B → Phi-4	8.6	6.8	8.0	12.4	6.2	4.4	3.2	26.6	+10.0	+6.8
Llama-3.1-8B → Qwen3-8B	7.6	5.0	17.2	9.0	8.2	5.4	5.6	35.4	+9.4	+3.8
Llama-3.2-1B → Llama-3.1-8B	5.8	3.4	21.6	5.2	3.2	2.0	2.4	33.2	+3.8	+1.4
Mistral-Nemo → Gemma-4-12B	8.8	9.6	10.2	11.8	5.6	4.4	3.8	32.4	+6.6	+2.8
Mistral-Nemo → Llama-3.1-8B	6.4	7.8	8.2	11.6	7.4	5.0	4.6	27.0	+8.8	+4.2
Mistral-Nemo → Phi-4	6.8	7.0	9.4	13.0	6.6	5.0	3.6	26.8	+11.0	+7.4
Mistral-Nemo → Qwen3-8B	7.8	8.4	19.0	9.6	6.2	5.6	6.4	41.6	+6.8	+0.4
Phi-4 → Gemma-4-12B	8.6	9.8	9.0	11.4	6.2	4.2	3.6	31.0	+5.8	+2.2
Phi-4 → Llama-3.1-8B	8.0	7.8	5.8	11.6	7.4	5.8	2.8	24.4	+9.6	+6.8
Phi-4 → Mistral-Nemo	9.0	7.6	9.0	10.0	7.8	4.8	3.0	28.6	+7.2	+4.2
Phi-4 → Qwen3-8B	8.0	9.2	18.2	8.6	7.8	5.2	5.4	40.8	+4.6	-0.8
Qwen3-8B → Gemma-4-12B	7.4	6.6	16.4	6.4	7.2	4.8	8.8	39.2	+4.6	-4.2
Qwen3-8B → Llama-3.1-8B	6.8	5.0	17.0	9.6	8.2	6.0	5.2	34.0	+10.6	+5.4
Qwen3-8B → Mistral-Nemo	7.8	4.6	17.0	7.8	6.0	6.0	8.2	37.6	+9.2	+1.0
Qwen3-8B → Phi-4	5.6	4.8	14.8	9.2	9.2	6.8	5.8	31.0	+11.2	+5.4
Mean	**7.3**	**6.9**	**12.3**	**10.3**	**7.1**	**5.0**	**4.7**	**31.3**	**+8.4**	**+3.6**

Table 11: Full error decomposition at n=1500 on NQ; alignment dataset: SQuAD. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	6.6	12.4	4.4	9.8	10.4	7.0	3.8	27.2	+4.4	+0.6
Gemma-4-12B → Mistral-Nemo	8.4	12.0	7.0	12.6	7.0	7.4	2.6	30.0	+8.0	+5.4
Gemma-4-12B → Phi-4	8.0	11.0	6.8	12.4	7.8	6.0	2.4	28.2	+7.4	+5.0
Gemma-4-12B → Qwen3-8B	7.8	11.8	14.4	7.8	7.2	7.8	7.8	41.8	+3.8	-4.0
Llama-3.1-8B → Gemma-4-12B	9.2	11.8	9.4	8.4	6.8	3.4	6.0	36.4	0.0	-6.0
Llama-3.1-8B → Mistral-Nemo	7.0	9.6	7.0	11.4	6.6	6.2	5.4	29.0	+8.0	+2.6
Llama-3.1-8B → Phi-4	8.0	11.6	7.8	13.0	5.2	4.6	4.2	31.6	+6.0	+1.8
Llama-3.1-8B → Qwen3-8B	9.0	12.6	18.4	7.6	7.0	4.2	6.8	46.8	-0.8	-7.6
Llama-3.2-1B → Llama-3.1-8B	6.4	6.6	21.6	4.6	3.0	2.0	2.6	37.2	0.0	-2.6
Mistral-Nemo → Gemma-4-12B	9.4	14.4	8.6	11.2	4.4	6.0	5.0	37.4	+2.8	-2.2
Mistral-Nemo → Llama-3.1-8B	6.6	11.8	8.2	11.4	8.4	5.0	3.6	30.2	+4.6	+1.0
Mistral-Nemo → Phi-4	7.6	11.8	9.6	12.2	6.4	4.8	3.8	32.8	+5.2	+1.4
Mistral-Nemo → Qwen3-8B	7.6	12.0	19.2	9.8	4.2	5.4	8.4	47.2	+3.2	-5.2
Phi-4 → Gemma-4-12B	7.8	9.6	9.0	12.2	5.6	4.2	4.2	30.6	+6.8	+2.6
Phi-4 → Llama-3.1-8B	8.4	12.6	6.2	11.2	7.6	5.4	2.6	29.8	+4.0	+1.4
Phi-4 → Mistral-Nemo	7.0	8.6	7.4	12.0	7.8	6.4	3.0	26.0	+9.8	+6.8
Phi-4 → Qwen3-8B	8.6	10.0	17.4	8.0	7.4	6.0	5.8	41.8	+4.0	-1.8
Qwen3-8B → Gemma-4-12B	7.2	9.0	15.4	6.6	7.0	5.8	9.0	40.6	+3.4	-5.6
Qwen3-8B → Llama-3.1-8B	8.0	7.4	17.2	8.4	7.4	5.8	6.0	38.6	+6.8	+0.8
Qwen3-8B → Mistral-Nemo	8.0	8.2	18.4	7.6	6.8	4.6	7.4	42.0	+4.0	-3.4
Qwen3-8B → Phi-4	8.6	10.0	15.2	6.2	7.8	6.4	7.2	41.0	+2.6	-4.6
Mean	**7.9**	**10.7**	**11.8**	**9.7**	**6.8**	**5.4**	**5.1**	**35.5**	**+4.5**	**-0.6**

Table 12: Full error decomposition at n=1500 on NQ; alignment dataset: TriviaQA. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	10.4	28.6	7.6	6.0	6.2	3.8	8.0	54.6	-18.8	-26.8
Gemma-4-12B → Mistral-Nemo	11.8	27.6	10.0	9.2	4.2	4.4	5.4	54.8	-14.0	-19.4
Gemma-4-12B → Phi-4	11.2	21.4	7.0	9.2	5.4	5.8	4.8	44.4	-6.4	-11.2
Gemma-4-12B → Qwen3-8B	5.8	19.0	7.2	9.8	10.2	15.0	4.8	36.8	+5.8	+1.0
Llama-3.1-8B → Gemma-4-12B	7.8	23.6	5.4	9.8	7.6	7.4	5.2	42.0	-6.4	-11.6
Llama-3.1-8B → Mistral-Nemo	9.2	23.0	7.2	9.2	6.0	6.0	6.0	45.4	-7.8	-13.8
Llama-3.1-8B → Phi-4	9.2	18.0	6.4	11.8	5.0	6.0	4.4	38.0	-0.2	-4.6
Llama-3.1-8B → Qwen3-8B	6.4	15.6	8.8	10.2	10.4	13.8	3.4	34.2	+8.4	+5.0
Llama-3.2-1B → Llama-3.1-8B	5.6	6.2	21.8	5.4	3.4	1.8	2.2	35.8	+1.0	-1.2
Mistral-Nemo → Gemma-4-12B	10.8	24.6	5.0	9.8	5.6	9.6	3.8	44.2	-5.2	-9.0
Mistral-Nemo → Llama-3.1-8B	11.4	27.4	6.0	6.6	6.4	7.2	5.6	50.4	-13.6	-19.2
Mistral-Nemo → Phi-4	10.4	22.2	7.0	9.4	6.2	7.4	4.0	43.6	-5.4	-9.4
Mistral-Nemo → Qwen3-8B	8.0	20.6	8.0	9.4	9.4	16.6	3.2	39.8	+5.4	+2.2
Phi-4 → Gemma-4-12B	9.2	19.4	8.4	10.8	6.0	4.8	3.8	40.8	-3.8	-7.6
Phi-4 → Llama-3.1-8B	9.0	19.0	5.4	10.6	6.2	6.2	4.0	37.4	-2.2	-6.2
Phi-4 → Mistral-Nemo	9.0	15.4	8.2	10.0	6.0	5.6	4.8	37.4	+0.2	-4.6
Phi-4 → Qwen3-8B	4.6	18.2	8.2	12.0	9.6	15.2	3.6	34.6	+9.0	+5.4
Qwen3-8B → Gemma-4-12B	8.4	14.0	17.8	5.4	6.4	3.4	9.6	49.8	-5.2	-14.8
Qwen3-8B → Llama-3.1-8B	9.6	11.6	18.2	6.8	5.8	4.8	7.6	47.0	0.0	-7.6
Qwen3-8B → Mistral-Nemo	9.6	7.6	17.0	6.0	7.0	6.0	7.2	41.4	+4.4	-2.8
Qwen3-8B → Phi-4	9.4	11.0	16.4	5.4	7.6	5.2	7.4	44.2	-0.4	-7.8
Mean	**8.9**	**18.8**	**9.9**	**8.7**	**6.7**	**7.4**	**5.2**	**42.7**	**-2.6**	**-7.8**

B.2.2 TRANSFER CURVES

Figures 3–5 show the complete AUROC curves across all 21 source–target pairs and all three evaluation datasets. For each evaluation dataset, we report both same-dataset and cross-dataset alignment.

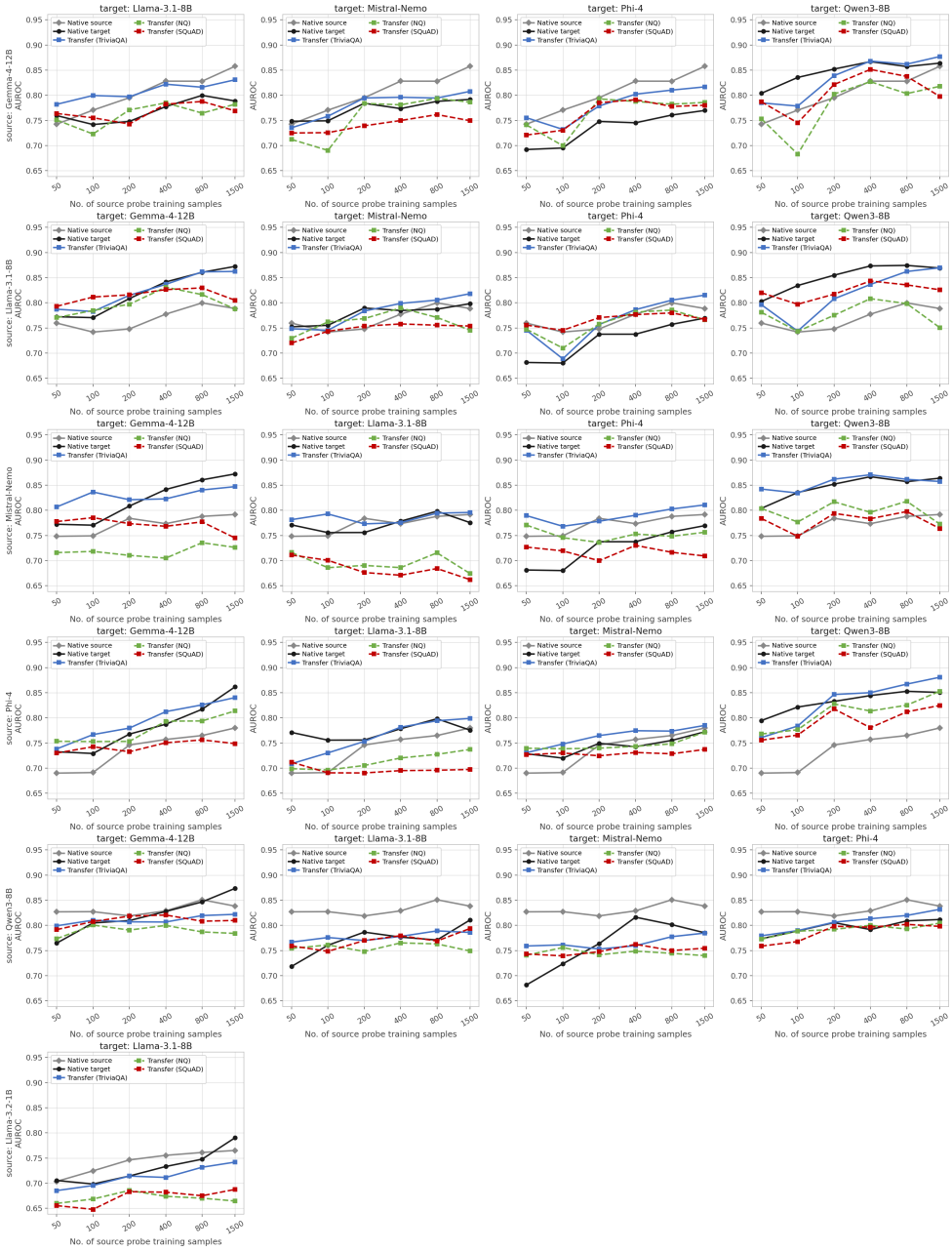

Figure 3: Probe transfer performance on TriviaQA as a function of the source probe training budget for all 21 source–target pairs. Native target is the target probe trained on target SE labels, while native source is the source probe evaluated against source SE labels. The transferred probes use alignment-based target layer selection without target SE labels. Alignment data are drawn from TriviaQA, NQ, or SQuAD. Same-dataset alignment uses TriviaQA representations; the other two curves use cross-dataset alignment. Affine Alignment uses $\lambda = 10^4$.

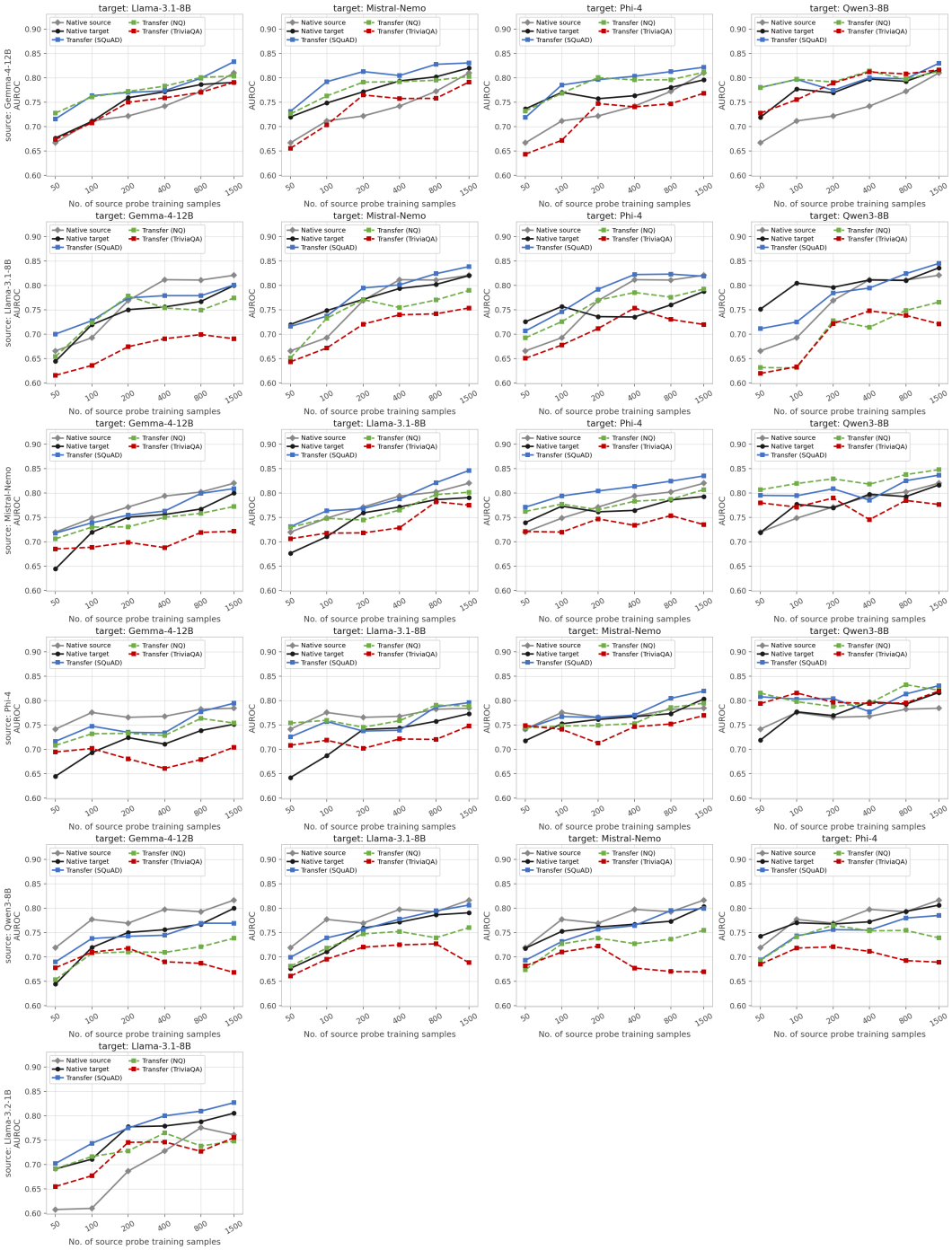

Figure 4: Probe transfer performance on SQuAD as a function of the source probe training budget for all 21 source–target pairs. Native target is the target probe trained on target SE labels, while native source is the source probe evaluated against source SE labels. Transferred probes use alignment-based target layer selection. Alignment data are drawn from SQuAD, NQ, or TriviaQA. Same-dataset alignment uses SQuAD representations; the other two curves use cross-dataset alignment. Affine Alignment uses $\lambda = 10^4$.

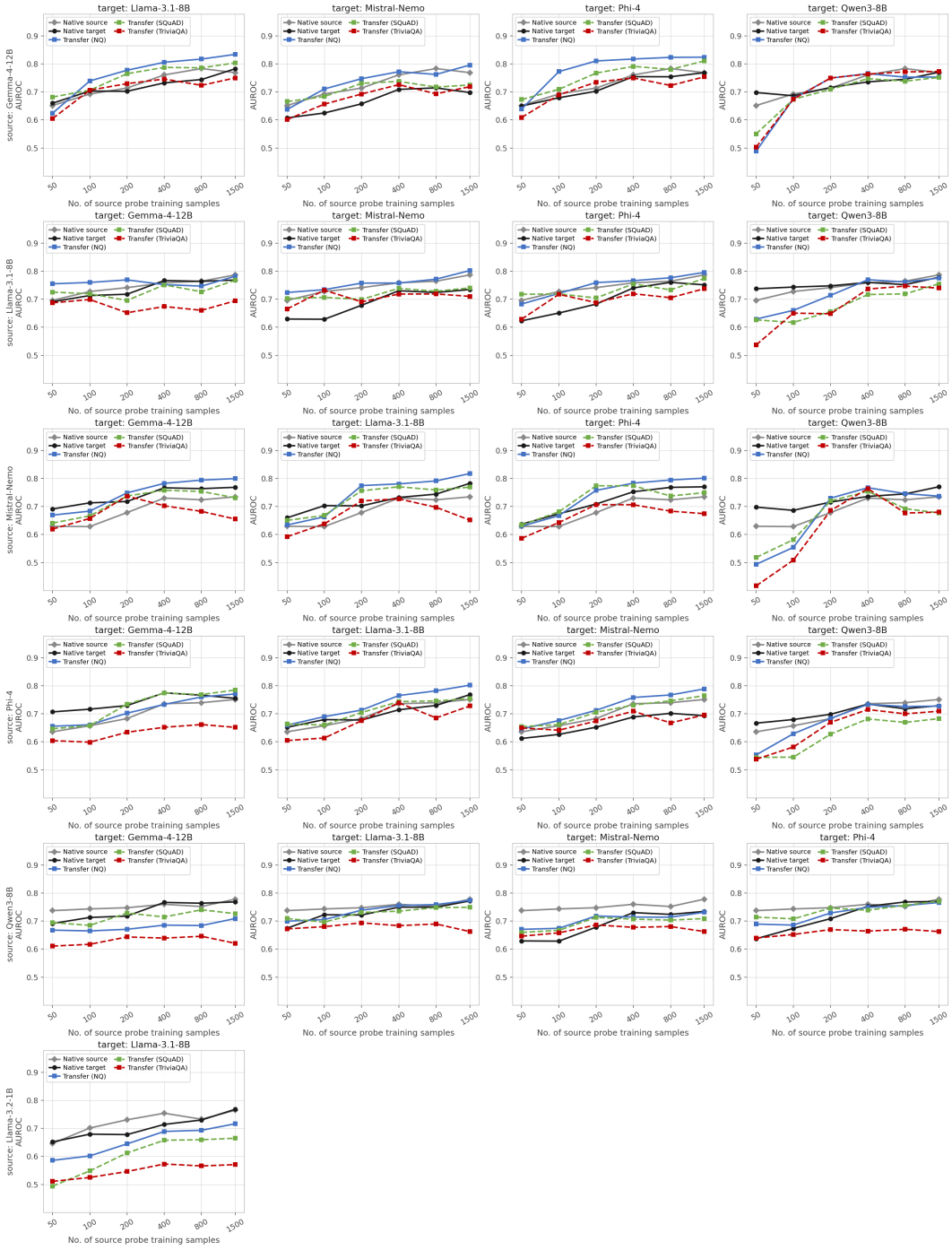

Figure 5: Probe transfer performance on Natural Questions (NQ) as a function of the source probe training budget for all 21 source–target pairs. Native target is the target probe trained on target SE labels, while native source is the source probe evaluated against source SE labels. Transferred probes use alignment-based target layer selection. Alignment data are drawn from NQ, TriviaQA, or SQuAD. Same-dataset alignment uses NQ representations; the other two curves use cross-dataset alignment. Affine Alignment uses $\lambda = 10^4$.

B.3 Alternative Target Layer Selection Strategies

We compare the target layer selection strategy used in the main experiments with the three alternatives introduced in Section 3.3: the final transformer layer, full-budget AUROC-based selection using the complete target SE label set, and low-budget AUROC-based selection using a small subsample of target SE labels. For all strategies, the source layer is selected by maximizing AUROC on the source SE labels.

For each alternative, we report the full error decomposition at $n = 1500$ and the corresponding probe-budget transfer curves under same-dataset alignment on TriviaQA. The native target baseline uses the target layer selected by the corresponding layer-selection strategy. Overall, the alternative strategies yield qualitatively similar transfer behavior to the alignment-based strategy used in the main experiments, although absolute performance varies across selection methods.

B.3.1 Final-Layer Selection

This strategy uses the final transformer layer for every target model and requires no target SE labels. Table 13 and Figure 6 report the corresponding results.

Table 13: Full error decomposition at n=1500 on TriviaQA; alignment dataset: TriviaQA, using final-layer target selection. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	6.0	11.4	6.4	5.2	4.8	5.0	3.0	26.8	-1.2	-4.2
Gemma-4-12B → Mistral-Nemo	4.4	9.4	8.0	7.2	5.2	5.6	2.2	24.0	+3.4	+1.2
Gemma-4-12B → Phi-4	4.6	8.0	10.8	5.8	6.2	5.2	2.4	25.8	+3.0	+0.6
Gemma-4-12B → Qwen3-8B	3.0	7.4	9.2	3.8	8.6	8.4	3.6	23.2	+4.8	+1.2
Llama-3.1-8B → Gemma-4-12B	6.4	8.4	4.6	8.4	6.6	4.4	3.6	23.0	+4.4	+0.8
Llama-3.1-8B → Mistral-Nemo	5.6	6.8	8.2	9.6	6.0	4.2	3.8	24.4	+7.0	+3.2
Llama-3.1-8B → Phi-4	7.0	6.0	8.6	8.8	7.0	5.2	2.2	23.8	+8.0	+5.8
Llama-3.1-8B → Qwen3-8B	4.2	6.2	9.4	8.8	9.2	7.2	2.8	22.6	+9.8	+7.0
Llama-3.2-1B → Llama-3.1-8B	8.8	8.0	29.2	7.8	2.0	2.8	5.4	51.4	+2.6	-2.8
Mistral-Nemo → Gemma-4-12B	6.8	8.6	6.4	8.4	6.6	4.6	3.4	25.2	+4.4	+1.0
Mistral-Nemo → Llama-3.1-8B	6.6	10.0	7.2	7.8	7.6	4.2	3.2	27.0	+2.0	-1.2
Mistral-Nemo → Phi-4	6.0	6.6	7.2	7.6	7.6	5.2	4.0	23.8	+6.2	+2.2
Mistral-Nemo → Qwen3-8B	3.6	6.2	9.6	7.4	10.2	7.4	4.0	23.4	+8.6	+4.6
Phi-4 → Gemma-4-12B	6.0	6.4	8.8	6.8	9.2	4.2	2.4	23.6	+4.6	+2.2
Phi-4 → Llama-3.1-8B	8.8	7.4	9.2	6.6	6.8	4.8	2.2	27.6	+4.0	+1.8
Phi-4 → Mistral-Nemo	7.2	5.8	9.0	7.2	7.8	5.0	2.2	24.2	+6.4	+4.2
Phi-4 → Qwen3-8B	4.8	7.8	5.6	6.0	10.0	5.6	3.6	21.8	+3.8	+0.2
Qwen3-8B → Gemma-4-12B	4.2	6.8	10.6	5.8	7.6	4.2	2.6	24.2	+3.2	+0.6
Qwen3-8B → Llama-3.1-8B	5.2	7.0	12.2	5.0	6.4	4.0	3.6	28.0	+2.0	-1.6
Qwen3-8B → Mistral-Nemo	4.6	5.6	13.4	5.6	6.8	4.2	3.2	26.8	+4.2	+1.0
Qwen3-8B → Phi-4	4.4	6.2	10.2	5.2	7.8	2.4	2.8	23.6	+1.4	-1.4
Mean	**5.6**	**7.4**	**9.7**	**6.9**	**7.1**	**4.9**	**3.2**	**25.9**	**+4.4**	**+1.3**

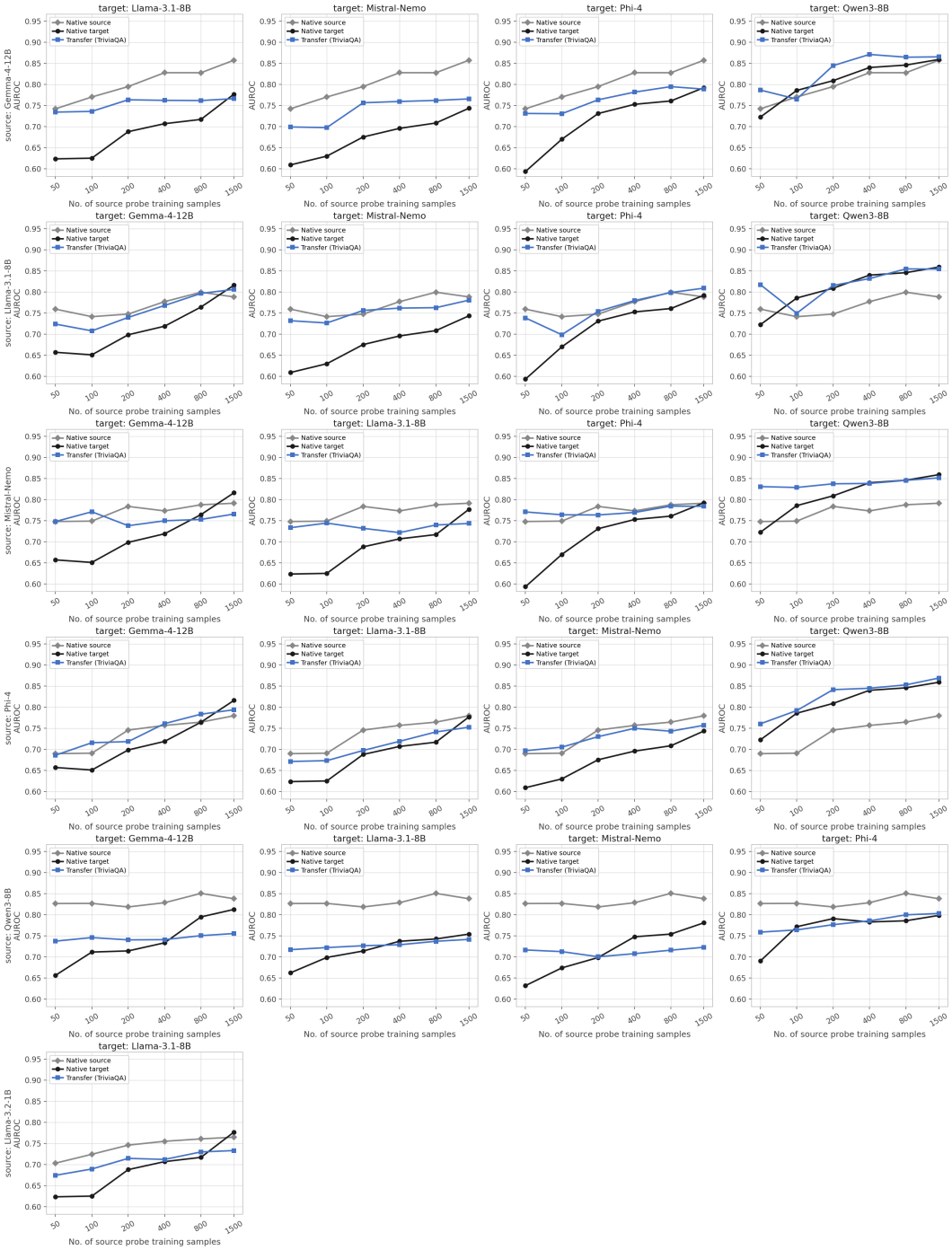

Figure 6: Probe transfer performance on TriviaQA as a function of the source probe training budget for all 21 source–target pairs using final-layer target selection. Native target is the target probe trained on target SE labels, while native source is the source probe evaluated against source SE labels. Same-dataset alignment uses TriviaQA representations. Affine Alignment uses $\lambda = 10^4$.

B.3.2 FULL-BUDGET AUROC-BASED SELECTION

This strategy selects the target layer that maximizes target probe AUROC using the full target SE label set. Table 14 and Figure 7 report the corresponding results.

Table 14: Full error decomposition at n=1500 on TriviaQA; alignment dataset: TriviaQA, using full-budget AUROC-based target-layer selection. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	4.8	9.4	7.8	6.4	5.2	3.6	2.6	24.6	+0.6	-2.0
Gemma-4-12B → Mistral-Nemo	5.0	6.6	8.4	6.6	4.8	5.2	2.6	22.6	+5.2	+2.6
Gemma-4-12B → Phi-4	4.0	6.2	9.8	6.4	6.4	6.2	2.2	22.2	+6.4	+4.2
Gemma-4-12B → Qwen3-8B	2.6	4.8	10.0	4.2	9.2	7.6	3.0	20.4	+7.0	+4.0
Llama-3.1-8B → Gemma-4-12B	6.2	5.8	6.6	8.6	6.2	2.4	4.0	22.6	+5.2	+1.2
Llama-3.1-8B → Mistral-Nemo	7.8	4.8	9.0	7.4	6.6	3.4	3.2	24.8	+6.0	+2.8
Llama-3.1-8B → Phi-4	6.8	5.2	9.4	9.0	6.6	4.4	2.6	24.0	+8.2	+5.6
Llama-3.1-8B → Qwen3-8B	4.2	3.6	9.8	8.8	9.8	6.8	2.2	19.8	+12.0	+9.8
Llama-3.2-1B → Llama-3.1-8B	8.2	6.6	28.6	8.4	2.0	3.4	5.4	48.8	+5.2	-0.2
Mistral-Nemo → Gemma-4-12B	6.8	8.0	7.4	8.4	7.6	3.6	2.4	24.6	+4.0	+1.6
Mistral-Nemo → Llama-3.1-8B	5.8	8.0	7.8	8.6	7.2	3.6	3.6	25.2	+4.2	+0.6
Mistral-Nemo → Phi-4	5.6	5.2	8.0	8.0	8.2	4.4	3.4	22.2	+7.2	+3.8
Mistral-Nemo → Qwen3-8B	3.6	3.8	10.0	7.4	11.8	7.0	2.4	19.8	+10.6	+8.2
Phi-4 → Gemma-4-12B	6.2	5.8	9.6	6.6	9.2	3.4	2.4	24.0	+4.2	+1.8
Phi-4 → Llama-3.1-8B	8.4	5.4	9.0	7.0	6.8	5.0	2.2	25.0	+6.6	+4.4
Phi-4 → Mistral-Nemo	8.0	5.2	10.2	6.4	7.4	3.8	2.6	26.0	+5.0	+2.4
Phi-4 → Qwen3-8B	4.4	3.4	6.6	6.4	11.2	4.6	2.4	16.8	+7.6	+5.2
Qwen3-8B → Gemma-4-12B	4.6	4.0	10.8	5.4	7.8	4.0	2.4	21.8	+5.4	+3.0
Qwen3-8B → Llama-3.1-8B	4.4	4.8	13.0	5.8	6.4	3.2	3.6	25.8	+4.2	+0.6
Qwen3-8B → Mistral-Nemo	4.8	2.8	12.2	5.4	6.4	5.4	3.6	23.4	+8.0	+4.4
Qwen3-8B → Phi-4	3.4	4.6	10.2	6.2	7.6	2.4	3.0	21.2	+4.0	+1.0
Mean	**5.5**	**5.4**	**10.2**	**7.0**	**7.4**	**4.4**	**2.9**	**24.1**	**+6.0**	**+3.1**

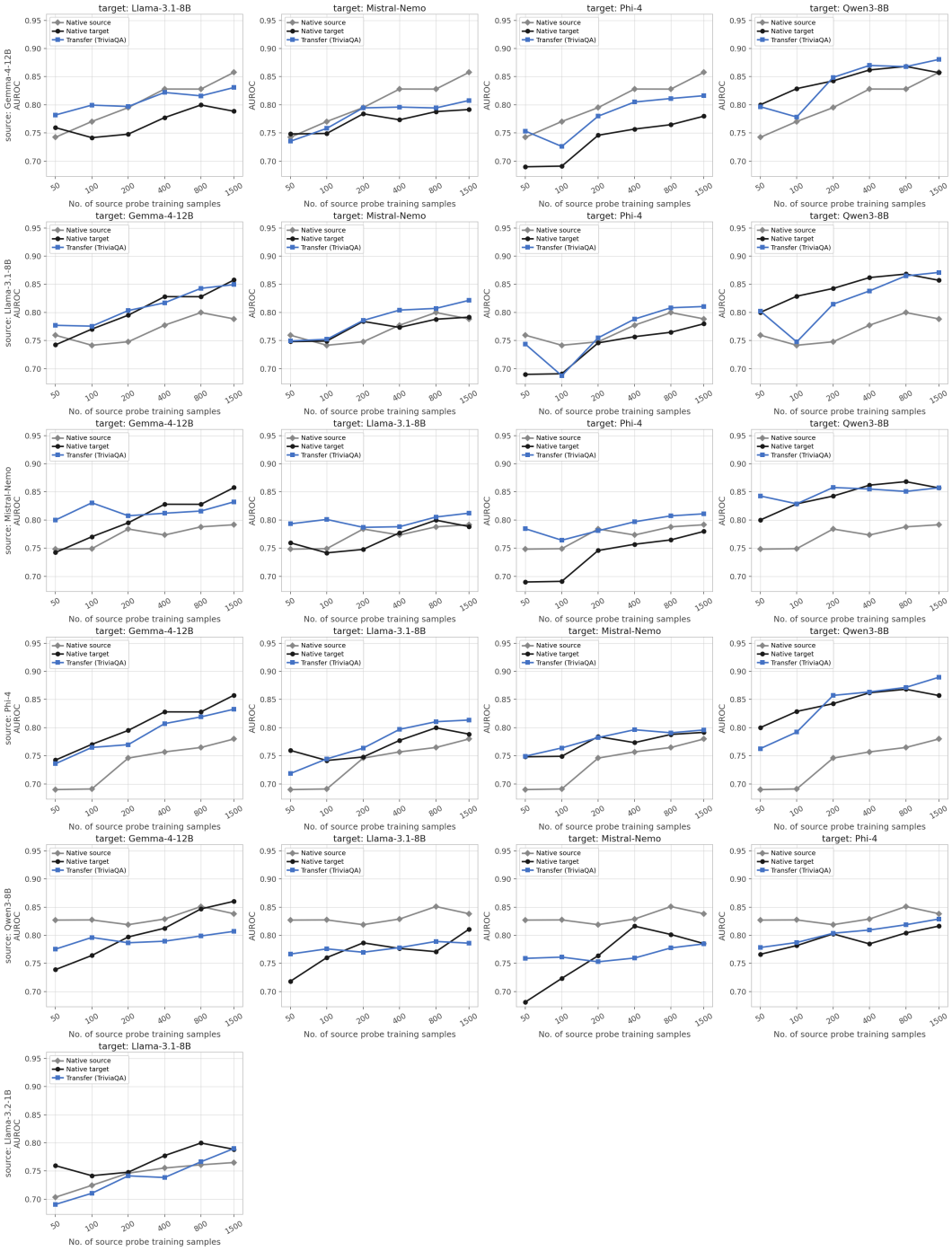

Figure 7: Probe transfer performance on TriviaQA as a function of the source probe training budget for all 21 source–target pairs using full-budget AUROC-based target layer selection. Native target is the target probe trained on target SE labels, while native source is the source probe evaluated against source SE labels. Same-dataset alignment uses TriviaQA representations. Affine Alignment uses $\lambda = 10^4$.

B.3.3 LOW-BUDGET AUROC-BASED SELECTION

This strategy performs AUROC-based target layer selection using a subsample of 150 target SE labels with cross-validation. Table 15 and Figure 8 report the corresponding results.

Table 15: Full error decomposition at n=1500 on TriviaQA; alignment dataset: TriviaQA, using low-budget AUROC-based target-layer selection. Entries are percentages of the 500 held-out test examples. Transfer error is the classification error of the transferred probe with respect to the target SE labels. $\Delta_{\text{excl}} = EF_{\text{only}} + FD_{\text{only}} - F_{\text{only}}$ is the corrective margin excluding the triple-interaction region, while $\Delta_{\text{acc}} = \Delta_{\text{excl}} - EFD$ is the exact change in target classification accuracy. The mean row reports the macro-average across the 21 source–target pairs.

Source → Target	E_{only}	F_{only}	D_{only}	EF_{only}	ED_{only}	FD_{only}	EFD	Transfer error	Δ_{excl}	Δ_{acc}
Gemma-4-12B → Llama-3.1-8B	4.8	9.4	7.8	6.4	5.2	3.6	2.6	24.6	+0.6	-2.0
Gemma-4-12B → Mistral-Nemo	4.8	6.4	8.6	6.8	5.2	5.0	2.2	22.0	+5.4	+3.2
Gemma-4-12B → Phi-4	4.2	5.8	9.6	6.2	6.0	6.4	2.6	22.2	+6.8	+4.2
Gemma-4-12B → Qwen3-8B	2.4	4.6	10.2	4.4	9.4	7.4	2.8	20.0	+7.2	+4.4
Llama-3.1-8B → Gemma-4-12B	6.2	5.8	6.6	8.6	6.2	2.4	4.0	22.6	+5.2	+1.2
Llama-3.1-8B → Mistral-Nemo	7.6	4.2	8.4	7.6	6.8	4.0	3.0	23.2	+7.4	+4.4
Llama-3.1-8B → Phi-4	6.8	4.8	9.4	9.0	7.0	4.4	2.2	23.2	+8.6	+6.4
Llama-3.1-8B → Qwen3-8B	4.8	4.2	8.6	8.2	10.0	8.0	2.0	19.6	+12.0	+10.0
Llama-3.2-1B → Llama-3.1-8B	8.2	6.6	28.6	8.4	2.0	3.4	5.4	48.8	+5.2	-0.2
Mistral-Nemo → Gemma-4-12B	6.8	8.0	7.4	8.4	7.6	3.6	2.4	24.6	+4.0	+1.6
Mistral-Nemo → Llama-3.1-8B	5.8	8.0	7.8	8.6	7.2	3.6	3.6	25.2	+4.2	+0.6
Mistral-Nemo → Phi-4	6.2	4.8	8.4	7.4	8.2	4.0	3.4	22.8	+6.6	+3.2
Mistral-Nemo → Qwen3-8B	3.8	3.8	9.4	7.2	11.4	7.6	2.8	19.8	+11.0	+8.2
Phi-4 → Gemma-4-12B	6.2	5.8	9.6	6.6	9.2	3.4	2.4	24.0	+4.2	+1.8
Phi-4 → Llama-3.1-8B	8.4	5.4	9.0	7.0	6.8	5.0	2.2	25.0	+6.6	+4.4
Phi-4 → Mistral-Nemo	7.4	4.6	9.8	7.0	7.4	4.2	2.6	24.4	+6.6	+4.0
Phi-4 → Qwen3-8B	4.4	3.2	6.2	6.4	10.8	5.0	2.8	16.6	+8.2	+5.4
Qwen3-8B → Gemma-4-12B	4.6	4.0	10.8	5.4	7.8	4.0	2.4	21.8	+5.4	+3.0
Qwen3-8B → Llama-3.1-8B	4.4	4.8	13.0	5.8	6.4	3.2	3.6	25.8	+4.2	+0.6
Qwen3-8B → Mistral-Nemo	5.2	4.0	13.4	5.0	6.0	4.2	4.0	26.6	+5.2	+1.2
Qwen3-8B → Phi-4	3.0	4.6	10.4	6.6	7.4	2.2	3.2	21.2	+4.2	+1.0
Mean	**5.5**	**5.4**	**10.1**	**7.0**	**7.3**	**4.5**	**3.0**	**24.0**	**+6.1**	**+3.2**

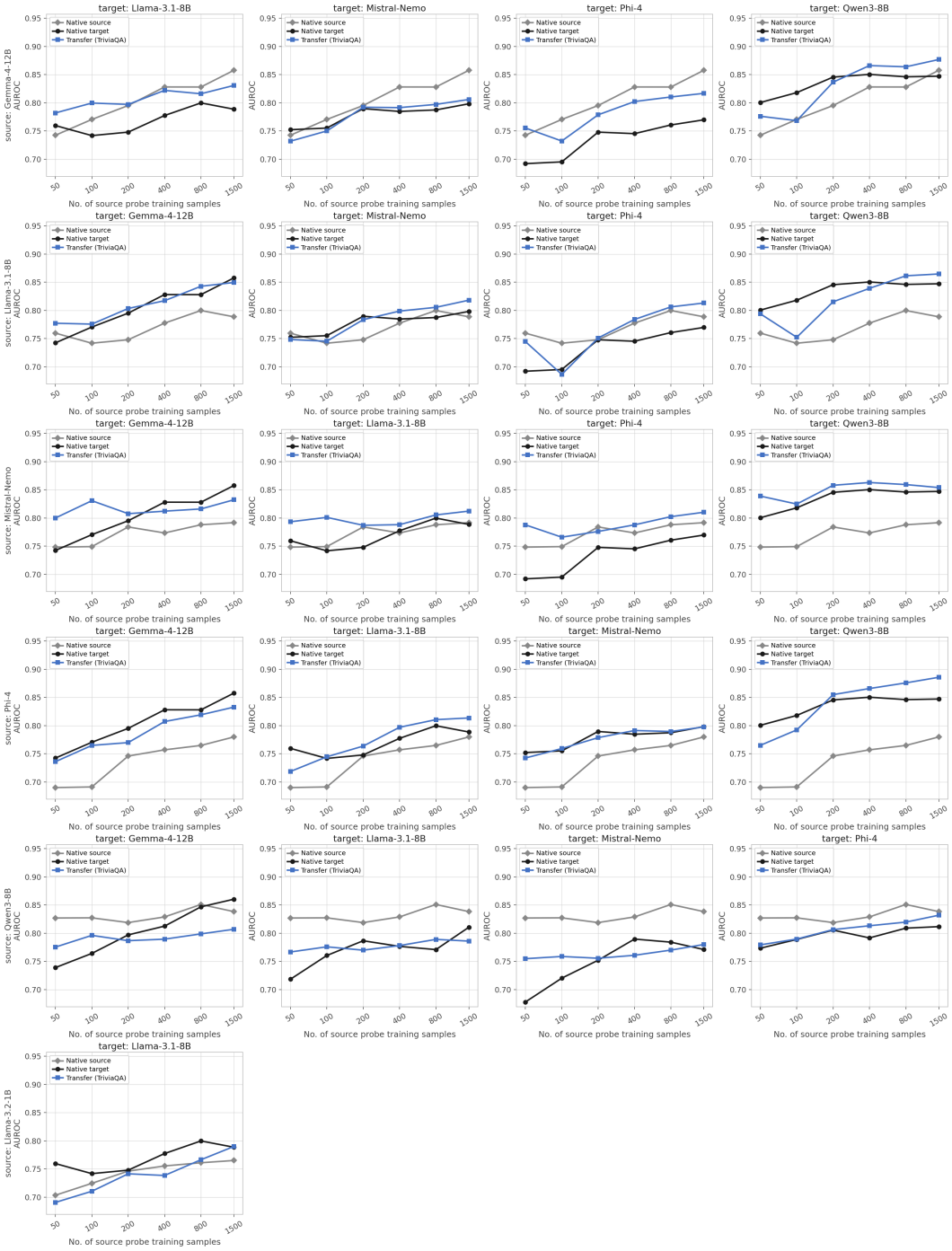

Figure 8: Probe transfer performance on TriviaQA as a function of the source probe training budget for all 21 source–target pairs using low-budget AUROC-based target layer selection. Native target is the target probe trained on target SE labels, while native source is the source probe evaluated against source SE labels. Same-dataset alignment uses TriviaQA representations. Affine Alignment uses $\lambda = 10^4$.