

# When In-Distribution Gains Fail: Evaluating Weak-to-Strong Reward Models under Preference Shift

Khoi Le<sup>†\*1</sup> Tri Cao<sup>\*1</sup> Phong Nguyen<sup>2</sup> Cong-Duy Nguyen<sup>2</sup>,  
Anh-Tuan Luu<sup>3,2</sup> Miao Chunyan<sup>3</sup> See-Kiong Ng<sup>1</sup> Thong Nguyen<sup>1</sup>

<sup>1</sup> National University of Singapore, <sup>2</sup> VinUniversity, <sup>3</sup> Nanyang Technological University

## Abstract

Weak-to-strong (W2S) generalization is a promising framework for scalable oversight, yet existing evaluations often test students under matched train–test distributions. Therefore, we study W2S preference learning under zero-shot distribution shift and find that strong students trained on weak preference labels can appear successful in-distribution while failing to transfer across preference datasets. We provide evidence consistent with a representational failure mode in which weak-supervised fine-tuning may move the strong model toward source-domain features instead of maintaining broadly transferable preference representations. To mitigate this, we propose *Representation Anchoring* (ANCHOR), a simple yet effective regularizer that constrains excessive drift from the pretrained strong model’s representation space during fine-tuning, while still allowing task-relevant adaptation. Across preference domains, datasets, and model families, ANCHOR improves out-of-distribution transfer while maintaining competitive in-distribution performance. Together, our evaluation protocol, transfer-aware metrics, and method expose hidden brittleness in current W2S reward modeling and provide a practical path toward more robust preference transfer.<sup>1</sup>

## 1 Introduction

As AI systems become increasingly capable, a central challenge is how to provide reliable supervision when humans cannot consistently evaluate model outputs due to limited expertise, cost, or scale. This challenge has motivated work on scalable oversight, which studies how weaker supervisors can guide stronger systems when direct supervision is costly, unreliable, or incomplete (Amodei et al., 2016; Leike et al., 2018; Bowman et al., 2022).

<sup>1</sup>Code: [https://github.com/TokisakiKurumi2001/w2s\\_ood](https://github.com/TokisakiKurumi2001/w2s_ood)

\*Contributes equally, <sup>†</sup>Corresponding author.  
Email: minhkhoi.le@u.nus.edu

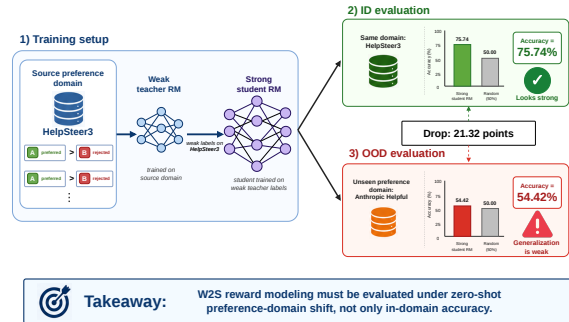


Figure 1: **ID success can hide OOD failure.** Although the resulting student can perform well on in-domain evaluation, its performance can degrade substantially when tested on unseen preference domains. This motivates evaluating W2S reward models under zero-shot preference-domain shift rather than relying only on in-domain accuracy.

Weak-to-strong generalization (W2S) provides a concrete abstraction of this problem: a stronger student is trained on labels produced by a weaker teacher, with the goal of eliciting capabilities beyond the teacher’s own performance (Burns et al., 2023).

Recent W2S methods have shown encouraging results across reasoning, mathematical problem solving, instruction tuning, and preference optimization (Burns et al., 2023; Yang et al., 2024; Li et al., 2024d; Zhu et al., 2025). These findings suggest that strong students can often extract useful task structure from imperfect weak supervision, enabling them to outperform their weak teachers rather than simply inherit the teacher’s limitations.

However, most existing W2S evaluations still test students under matched or near-matched train–test distributions, which may overestimate alignment reliability in realistic deployment. We instead study preference-domain shift: the weak supervisor and student are trained on one preference dataset and evaluated zero-shot on another dataset within the same broad preference category, where prompts,

response styles, annotation protocols, and preference policies may differ. Unlike an ordinary held-out split, this setting tests whether weak-supervised reward models learn transferable preference criteria rather than dataset-specific patterns. This gap motivates our central question: *does weak-to-strong preference learning transfer alignment behavior beyond the distribution from which weak supervision is generated?*

To answer this question, we introduce a zero-shot distribution-shift evaluation protocol for W2S, focusing on reward modeling as a central component of preference-based alignment pipelines (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). Within each broad preference domain, such as helpfulness or harmlessness, models are trained on one dataset and evaluated both on its held-out test split and on other datasets from the same domain but with different annotation sources and preference styles. Across domains, datasets, and model families, we find that existing W2S methods often match or approach strong ground-truth performance in-distribution, yet remain fragile under cross-distribution evaluation. As shown in Figure 1, a reward model trained on HelpSteer3 (Wang et al., 2025) achieves strong in-distribution performance, but its accuracy drops substantially when evaluated on Anthropic Helpful (Bai et al., 2022).

To improve robustness under this shift, we propose *Representation Anchoring* (ANCHOR), which constrains excessive drift from the pretrained strong model’s representation space during fine-tuning. ANCHOR is designed to limit excessive distortion of broadly useful pretrained features while still allowing the reward model to adapt to weak preference labels. Across preference domains and model families, ANCHOR achieves the most balanced overall performance over existing W2S methods under both in-distribution and out-of-distribution evaluation.

In summary, our contributions are threefold:

- We introduce a zero-shot preference-domain shift protocol for evaluating W2S reward modeling across datasets, domains, and model families.
- Using this protocol, we show that existing W2S methods can appear successful in-distribution yet remain fragile under cross-domain evaluation, revealing that in-distribution performance can overestimate alignment reliability.
- We propose *Representation Anchoring* (ANCHOR), a simple yet effective method that regularizes the student toward the pretrained strong model’s representations, limiting excessive representation drift during weak-supervised training. Across empirical experiments, ANCHOR yields the most balanced gains under both in-distribution and out-of-distribution evaluation.

## 2 Preliminaries

### 2.1 Weak-to-Strong Generalization

Weak-to-strong (W2S) generalization studies whether a stronger model can be trained from supervision produced by a weaker model (Burns et al., 2023). In each source domain  $S$ , we instantiate three models: a weak supervisor  $w_S$ , a strong student  $m_S$ , and a strong ceiling model  $m_S^{\text{gt}}$ . The weak supervisor is trained on gold labels and then used to supervise the strong student. The strong ceiling model has the same base capacity as  $m_S$ , but is trained directly on gold labels and is used only for evaluation.

In this work, we study W2S in the setting of preference reward modeling, where the supervised object is a scalar reward model trained from pairwise preferences.

### 2.2 Preference Reward Modeling

A preference dataset consists of prompts paired with a preferred and rejected response,

$$\mathcal{D} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^N.$$

A reward model  $r_\theta(x, y) \in \mathbb{R}$  assigns a scalar score to each prompt–response pair. Following the Bradley–Terry model (Bradley and Terry, 1952), the probability of preferring  $y_i^+$  over  $y_i^-$  is

$$p_\theta(y_i^+ \succ y_i^- \mid x_i) = \sigma(r_\theta(x_i, y_i^+) - r_\theta(x_i, y_i^-)),$$

and the standard reward-modeling objective is

$$\mathcal{L}_{\text{RM}} = -\frac{1}{N} \sum_{i=1}^N \log \sigma(r_\theta(x_i, y_i^+) - r_\theta(x_i, y_i^-)).$$

In supervised reward modeling,  $(y_i^+, y_i^-)$  is determined by human preference labels, as in RLHF-style alignment pipelines (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). In W2S reward modeling, the strong student instead learns from preference labels induced by the weak supervisor.

### 2.3 Weak-to-Strong Reward-Modeling Protocol

Given a source training set  $S_{\text{train}}$ , we split it into disjoint supervisor and student subsets. We train  $w_S$  on gold preference labels from the supervisor subset, and then use  $w_S$  to annotate each example in the student subset with a soft preference label

$$q_i = p_{w_S}(y_i^+ \succ y_i^- \mid x_i).$$

The strong student  $m_S$  is trained only on these weak labels, while the strong ceiling model  $m_S^{\text{gt}}$  is trained on gold labels using the same strong model base capacity.

## 3 Problem definition

We now formalize the central question of this work: whether weak-to-strong reward modeling learns domain-general preference representations or merely fits source-domain preference patterns.

### 3.1 Zero-Shot Preference-Domain Shift

Let  $\mathcal{C}$  denote a broad preference category, such as helpfulness or harmlessness. Each category contains a collection of preference domains

$$\mathcal{K}_{\mathcal{C}} = \{k_1, k_2, \dots, k_M\},$$

where each  $k_i$  corresponds to a distinct data distribution with its own prompts, response styles, annotation protocol, and dataset artifacts. For each source domain  $k_i \in \mathcal{K}_{\mathcal{C}}$ , we train a weak-to-strong reward model  $r_{\theta}$  using only  $k_i^{\text{train}}$ . We evaluate  $r_{\theta}$  in-distribution on the held-out test set from the same domain,

$$\mathcal{E}_{\text{ID}}(i) = \{k_i^{\text{test}}\},$$

and out-of-distribution on the test sets from all remaining domains,

$$\mathcal{E}_{\text{OOD}}(i) = \{k_j^{\text{test}} : k_j \in \mathcal{K}_{\mathcal{C}}, j \neq i\}.$$

We focus on *zero-shot preference-domain transfer*: no target-domain examples are used to train the weak supervisor, generate weak labels, or tune the reward head. Hence, target-domain performance reflects transfer from source-domain weak supervision rather than target-domain adaptation.

This formulation differs from ordinary held-out evaluation. A held-out source test set asks whether the student learned the source preference distribution, while a target-domain test set asks whether the

learned preference signal remains useful when the prompt distribution, response style, or dataset artifacts change. This distinction is central to scalable oversight: weak supervision may be available only for limited behavioral slices, whereas deployed reward models must generalize across many shifted preference domains.

### 3.2 Transfer-Aware Metrics

Raw preference accuracy is necessary but insufficient for evaluating weak-to-strong reward modeling under preference-domain shift. A method may improve on the source domain while failing to transfer, or improve on target domains only by degrading source-domain behavior. We therefore report three transfer-aware metrics that separately capture source-domain improvement, target-domain transfer, and transfer after accounting for source-domain regression.

Let  $S \in \mathcal{K}_{\mathcal{C}}$  be the source preference domain used to train the weak supervisor  $w_S$  and the strong student  $m_S$ . For any evaluation domain  $E \in \mathcal{K}_{\mathcal{C}}$ , define

$$A_w(S, E) = \text{Acc}(w_S, E_{\text{test}}),$$

$$A_m(S, E) = \text{Acc}(m_S, E_{\text{test}}),$$

where  $\text{Acc}(\cdot, E_{\text{test}})$  is pairwise preference accuracy on the test split of domain  $E$ . The case  $E = S$  corresponds to in-distribution evaluation, while  $E = T \neq S$  corresponds to zero-shot out-of-distribution evaluation on a target domain  $T$ .

**Weak-to-Strong Raw Gain.** Weak-to-Strong Raw Gain (WRG) (Charikar et al., 2024) measures whether the strong student improves over the weak supervisor on the source domain:

$$\text{WRG}(S; m) = A_m(S, S) - A_w(S, S).$$

Positive WRG indicates successful in-distribution weak-to-strong improvement.

**Absolute OOD Gain.** For an unseen target domain  $T \neq S$ , Absolute OOD Gain (AOG) measures whether the weak-to-strong improvement transfers beyond the source domain:

$$\text{AOG}(S, T; m) = A_m(S, T) - A_w(S, T).$$

Positive AOG indicates that the strong student outperforms the weak supervisor under zero-shot preference-domain shift.

**Net Transfer Score.** OOD improvement is desirable only if it does not come at the cost of source-domain collapse. We therefore define the in-distribution regression cost as

$$C_{\text{ID}}(S; m) = \max(0, A_w(S, S) - A_m(S, S)),$$

and define the Net Transfer Score (NTS) as

$$\text{NTS}(S, T; m) = \text{AOG}(S, T; m) - C_{\text{ID}}(S; m).$$

When the strong student preserves or improves source-domain performance relative to the weak supervisor,  $C_{\text{ID}}(S; m) = 0$ , and NTS equals AOG. When target-domain gains are obtained by falling below the weak supervisor on the source domain, NTS penalizes this regression. Thus, NTS captures the central requirement of zero-shot weak-to-strong transfer: improving target-domain performance without sacrificing source-domain reliability.

## 4 Methodology

We now describe ANCHOR, our weak-to-strong method for zero-shot preference-domain transfer. As shown in Figure 2, ANCHOR augments standard reward-model training with a representation anchoring regularizer. The student is trained to match weak teacher preferences through the usual scalar reward modeling loss, while its response-token hidden states are simultaneously constrained to remain close to those of a frozen pretrained reference model. The final objective combines the preference loss ( $\mathcal{L}_{w2s}$ ) and the anchoring loss ( $\mathcal{L}_{\text{anchor}}$ ), allowing the student to learn from weak supervision while discouraging excessive deviation from pretrained representations that may support transfer.

### 4.1 Weak Preference Learning Objective

Let  $\mathcal{P}_S = \{(x_i, y_i^a, y_i^b)\}_{i=1}^N$  denote a set of source-domain prompt–response pairs, where  $x_i$  is a prompt and  $y_i^a, y_i^b$  are two candidate responses. In the weak-to-strong setting, the strong student does not observe the ground-truth preference label. Instead, a weak teacher reward model  $w_S$ , trained on source domain  $S$ , provides the soft preference target

$$q_i = \sigma(w_S(x_i, y_i^a) - w_S(x_i, y_i^b)),$$

where  $q_i$  is the teacher-assigned probability that  $y_i^a$  is preferred to  $y_i^b$ .

The strong student  $m_S$  predicts

$$p_i = \sigma(m_S(x_i, y_i^a) - m_S(x_i, y_i^b)).$$

Following the soft-label W2S objective of Burns et al. (2023), let  $z_i \sim \text{Bernoulli}(q_i)$  denote the preference label induced by the weak teacher. The expected negative log-likelihood under this teacher distribution is

$$\mathbb{E}_{z_i}[-\log p_\theta(z_i)] = -q_i \log p_i - (1 - q_i) \log(1 - p_i).$$

Averaging over the training examples gives

$$\mathcal{L}_{w2s} = -\frac{1}{N} \sum_{i=1}^N [q_i \log p_i + (1 - q_i) \log(1 - p_i)].$$

$\mathcal{L}_{w2s}$  trains the strong student to match the weak teacher’s soft preference distribution while retaining the standard scalar reward-modeling form, in which each prompt–response pair receives an independent score and pairwise preferences are derived from score differences. In ANCHOR, this objective provides the preference-learning signal and is combined with the representation-anchoring regularizer to encourage zero-shot transfer across preference domains.

### 4.2 Representation Anchoring Regularizer

Our central hypothesis is that the pretrained strong model contains broadly useful representations for preference judgment, while weak-supervised fine-tuning on a single source domain may move these representations toward source-specific features. ANCHOR addresses this by adding a frozen copy of the pretrained strong model as a training-time reference. Let  $m_S$  denote the trainable strong reward model, and let  $m_{\text{ref}}$  denote a frozen copy initialized from the same pretrained checkpoint before weak-supervised fine-tuning. The reference model is never updated and is discarded after training.

For a prompt–response pair  $(x, y)$ , we denote

$$H_\theta^\ell(x, y) \in \mathbb{R}^{T \times d}, \quad H_{\text{ref}}^\ell(x, y) \in \mathbb{R}^{T \times d}$$

as the hidden states at layer  $\ell$  from the student and reference models, where  $T$  is the sequence length and  $d$  is the hidden dimension. We use a response-token mask

$$m(x, y) \in \{0, 1\}^T$$

to exclude prompt tokens and padding tokens from the anchoring loss. Following that, the masked

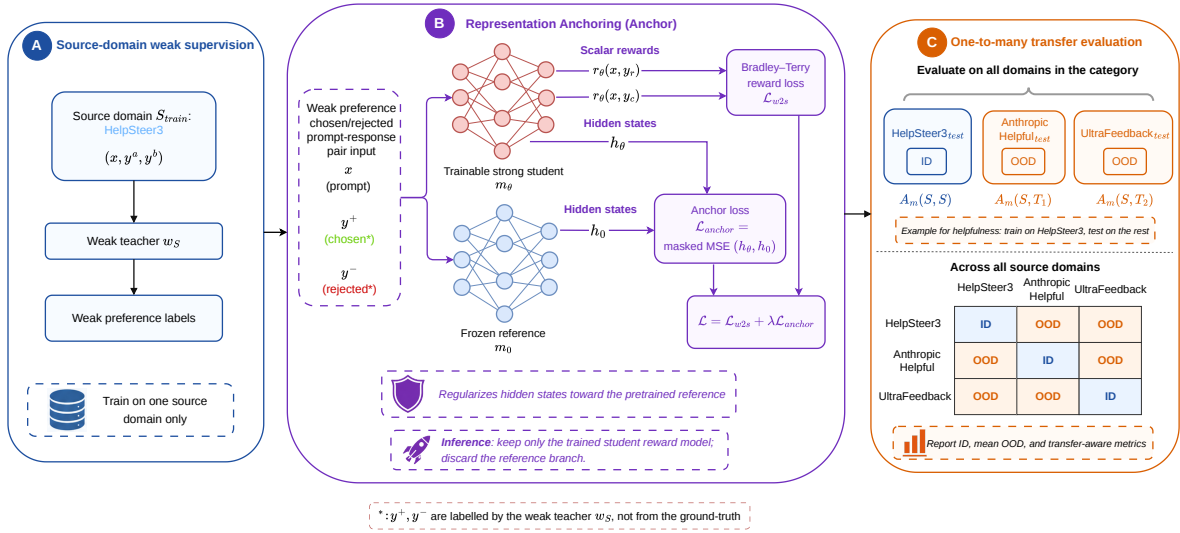


Figure 2: **Overview of ANCHOR under zero-shot preference-domain shift.** ANCHOR trains a strong reward model with the standard weak-to-strong preference loss  $\mathcal{L}_{w2s}$ , while regularizing its response-token hidden states toward a frozen pretrained reference model through  $\mathcal{L}_{\text{anchor}}$ . At inference, the reference model is discarded and only the learned scalar reward model is used for both in-distribution and zero-shot out-of-distribution evaluation.

hidden-state distance at layer  $\ell$  is defined as

$$D_{\text{anchor}}^{\ell}(x, y) = \frac{\sum_{t=1}^T m_t(x, y) \|\Delta_t^{\ell}(x, y)\|_2^2}{d \sum_{t=1}^T m_t(x, y)},$$

where

$$\Delta_t^{\ell}(x, y) = H_{\theta,t}^{\ell}(x, y) - H_{\text{ref},t}^{\ell}(x, y).$$

For the  $i$ -th pseudo-preference pair, the anchoring loss is computed on both candidate responses:

$$\mathcal{L}_{\text{anchor}}^{\ell}(i) = D_{\text{anchor}}^{\ell}(x_i, y_i^a) + D_{\text{anchor}}^{\ell}(x_i, y_i^b).$$

Therefore, the full ANCHOR objective on the  $i$ -th sample is

$$\mathcal{L}(i) = \mathcal{L}_{w2s}(i) + \lambda \mathcal{L}_{\text{anchor}}^{\ell}(i),$$

where  $\lambda$  controls the strength of representation preservation.

The two terms play complementary roles. The weak preference loss provides the task signal by training  $m_S$  to reproduce the weak teacher’s soft preferences, whereas the anchoring loss regularizes the internal response representations used to support these predictions. By regularizing the student’s representation space close to the pretrained reference space during source-domain fine-tuning, ANCHOR aims to preserve broadly useful pretrained features while still allowing the reward model to adapt to weak supervision.

### 4.3 ANCHOR Variants

We consider two variants of ANCHOR that differ in which layer representations are regularized.

**Last-Layer ANCHOR.** Our default variant applies the anchoring loss to the final transformer layer:

$$\mathcal{L}_{\text{Last}}(i) = \mathcal{L}_{\text{anchor}}^L(i),$$

where  $L$  is the number of transformer layers. This variant directly regularizes the representation used by the reward head and requires storing hidden states from only one layer. Because of its lower memory and computation cost, we use Last-Layer ANCHOR as the main version of our method.

**Middle-Layer ANCHOR.** We also evaluate a middle-layer variant, which anchors intermediate representations while leaving the final reward-facing layers more flexible. Let  $\mathcal{O}$  be a small set of normalized layer positions, for example

$$\mathcal{O} = \{\lfloor 0.5L \rfloor, \lfloor 0.75L \rfloor\}.$$

The middle-layer anchoring loss is

$$\mathcal{L}_{\text{MLA}}(i) = \frac{1}{|\mathcal{O}|} \sum_{\ell \in \mathcal{O}} \mathcal{L}_{\text{anchor}}^{\ell}(i).$$

The corresponding training objective now becomes

$$\mathcal{L}(i) = \mathcal{L}_{w2s}(i) + \lambda \mathcal{L}_{\text{MLA}}(i).$$

Middle-Layer ANCHOR is useful both as an alternative regularizer and as a diagnostic. If preserving intermediate representations improves OOD transfer, it provides evidence consistent with a role for representation preservation in weak-to-strong transfer. In our main experiments, we use Last-Layer ANCHOR for efficiency and compare it with Middle-Layer ANCHOR in the ablation study.

## 5 Experiment

### 5.1 Experimental Setup

**Datasets.** We evaluate on two preference categories. For **Helpful**, we use Anthropic Helpful from HH-RLHF (Bai et al., 2022), HelpSteer3-Preference (Wang et al., 2025), and UltraFeedback (Cui et al., 2024). For **Harmless**, we use Anthropic Harmless from HH-RLHF (Bai et al., 2022), PKU-SafeRLHF (Ji et al., 2025), and RAIL (Verma et al., 2025). When a dataset does not provide a validation split, we reserve 10% of its training split for validation. We then split the remaining training data into two equal parts: a gold-labeled subset  $\mathcal{D}_S^{\text{gold}}$ , used to train the weak teacher  $w_S$ , and a weak-to-strong subset  $\mathcal{D}_S^{\text{w2s}}$ , which is relabeled by  $w_S$  and used to train the strong student  $m_S$ .

**Models.** We report results on two model families. For Llama, we use Llama-3.2-1B-Instruct as the weak teacher and Llama-3.1-8B-Instruct as the strong student (Grattafiori et al., 2024). For Qwen, we use Qwen3-1.7B as the weak teacher and Qwen3-8B as the strong student (Yang et al., 2025a). All models are trained as scalar reward models using LoRA with rank = 16. Details of the hyperparameters can be found in Appendix A.

**Baselines.** We compare ANCHOR with three W2S preference-learning baselines:

- **Naive W2S** (Burns et al., 2023): trains the strong student directly on weak teacher labels using the standard weak preference loss  $\mathcal{L}_{\text{w2s}}$  from Section 4.1.
- **Confidence-based W2S** (Burns et al., 2023): augments naive W2S with an auxiliary confidence loss that encourages confident student predictions and reduces direct imitation of weak-teacher errors.
- **SEAM** (Li et al., 2025): uses the pretrained strong model to generate principle-based annotation rationales, while the weak model selects the final weak annotation.

**Metrics and reporting.** We use pairwise preference accuracy as the base measure and report the full accuracy table in Appendix D. In the main paper, we summarize performance using PGR (Burns et al., 2023) and three transfer-aware metrics: WRG, AOG, and NTS, all computed from mean accuracies over three random seeds.

### 5.2 Main results

Table 1 shows that in-distribution weak-to-strong gains do not always imply robust transfer to unseen preference domains. Naive and confidence-based training exhibit a similar pattern: they can achieve strong in-distribution performance, but their OOD transfer remains unstable. For example, in the harmless category with the Qwen model family, when training on Anthropic Harmless, Naive obtains tied-best in-distribution WRG score of 1.48, but transfers weakly to RAIL with only 0.52 in AOG. Similarly, when training on PKU-SafeRLHF, confidence-based filtering achieves a high in-distribution PGR of 1537.5%, but gives negative transfer to Anthropic Harmless with  $-0.39$  in NTS. These results suggest that improving or filtering weak source-domain supervision alone does not reliably produce transferable reward models.

SEAM shows a sharp preservation transfer trade-off: it can achieve competitive OOD AOG, but often at the cost of severe in-distribution degradation. In the helpful Llama setting, training on HelpSteer3 gives SEAM strong OOD AOG on Anthropic Helpful and UltraFeedback (9.95 and 6.25), yet its ID WRG drops to  $-15.67$ ; the same pattern appears for Qwen trained on Anthropic Helpful, with 5.49 OOD AOG on UltraFeedback but  $-1315.4\%$  ID PGR. In contrast, ANCHOR better preserves source-domain performance while improving transfer, raising AOG from 0.60 to 4.41 on Anthropic Helpful and NTS from 5.79 to 6.01 on HelpSteer3 when trained on UltraFeedback. Across Table 1, ANCHOR is best or tied in 41 of 60 WRG/AOG/NTS comparisons, supporting its stronger overall balance across ID and OOD evaluation rather than uniform dominance in every setting.

Nevertheless, ANCHOR does not guarantee positive transfer for every source-target pair. In particular, Qwen trained on HelpSteer3 retains a positive ID gain (WRG = 0.33) and transfers positively to UltraFeedback (NTS = 4.69), but remains below the weak teacher on Anthropic Helpful (NTS =  $-2.61$ ). This failure persists across

Table 1: **Weak-to-strong reward-model transfer under preference-domain shift.** We report in-distribution and zero-shot out-of-distribution transfer results across helpful and harmless preference domains.

| Helpful category                                            |              |             |              |              |             |             |              |             |             |             |             |             |              |              |             |             |             |             |
|-------------------------------------------------------------|--------------|-------------|--------------|--------------|-------------|-------------|--------------|-------------|-------------|-------------|-------------|-------------|--------------|--------------|-------------|-------------|-------------|-------------|
| <i>Llama-3.2-1B-Instruct</i> → <i>Llama-3.1-8B-Instruct</i> |              |             |              |              |             |             |              |             |             |             |             |             |              |              |             |             |             |             |
| Method                                                      | H3           |             |              |              |             |             | AH           |             |             |             |             |             | UF           |              |             |             |             |             |
|                                                             | In-dist.     |             | AH           |              | UF          |             | In-dist.     |             | H3          |             | UF          |             | In-dist.     |              | H3          |             | AH          |             |
|                                                             | PGR (%)      | WRG         | AOG          | NTS          | AOG         | NTS         | PGR (%)      | WRG         | AOG         | NTS         | AOG         | NTS         | PGR (%)      | WRG          | AOG         | NTS         | AOG         | NTS         |
| Naive                                                       | <u>32.23</u> | <u>1.71</u> | 1.46         | 1.46         | 4.8         | 4.8         | <u>98.7</u>  | <u>3.13</u> | <b>4.59</b> | <b>4.59</b> | <u>4.98</u> | <u>4.98</u> | <u>73.4</u>  | <u>1.74</u>  | <u>5.79</u> | <u>5.79</u> | <u>0.6</u>  | <u>0.6</u>  |
| Conf.-based                                                 | -39.9        | -2.12       | 8.06         | <u>5.94</u>  | <u>7.64</u> | <u>5.52</u> | 70.3         | 2.23        | <u>3.83</u> | <u>3.83</u> | 4.74        | 4.74        | 41.8         | 0.99         | 3.39        | 3.39        | 0.25        | 0.25        |
| SEAM                                                        | -295.1       | -15.67      | <b>9.95</b>  | -5.72        | 6.25        | -9.42       | -411.4       | -13.04      | -6.01       | -19.05      | -0.46       | -13.5       | -280.6       | -6.65        | -5.68       | -12.33      | 0.17        | -6.48       |
| ANCHOR                                                      | <b>67.2</b>  | <b>3.57</b> | <u>8.19</u>  | <b>8.19</b>  | <b>7.93</b> | <b>7.93</b> | <b>106.9</b> | <b>3.39</b> | 2.08        | 2.08        | <b>5.61</b> | <b>5.61</b> | <b>97.9</b>  | <b>2.32</b>  | <b>6.01</b> | <b>6.01</b> | <b>4.41</b> | <b>4.41</b> |
| <i>Qwen3-1.7B</i> → <i>Qwen3-8B</i>                         |              |             |              |              |             |             |              |             |             |             |             |             |              |              |             |             |             |             |
| Method                                                      | H3           |             |              |              |             |             | AH           |             |             |             |             |             | UF           |              |             |             |             |             |
|                                                             | In-dist.     |             | AH           |              | UF          |             | In-dist.     |             | H3          |             | UF          |             | In-dist.     |              | H3          |             | AH          |             |
|                                                             | PGR (%)      | WRG         | AOG          | NTS          | AOG         | NTS         | PGR (%)      | WRG         | AOG         | NTS         | AOG         | NTS         | PGR (%)      | WRG          | AOG         | NTS         | AOG         | NTS         |
| Naive                                                       | -87.7        | -3.06       | <u>-0.04</u> | <u>-3.1</u>  | 0.17        | -2.89       | 173.1        | 0.45        | 5.57        | 5.57        | 5.15        | <u>5.15</u> | -89.1        | -0.41        | <b>3.61</b> | <u>3.2</u>  | 1.37        | 0.96        |
| Conf.-based                                                 | -65.9        | -2.3        | -1.58        | -3.88        | 4.4         | 2.1         | <u>661.5</u> | <u>1.72</u> | 6.78        | 6.78        | 4.92        | 4.92        | <u>-39.1</u> | -0.18        | 1.42        | 1.24        | <u>3.26</u> | 3.08        |
| SEAM                                                        | -332.1       | -11.59      | <b>8.2</b>   | -3.39        | <b>5.79</b> | -5.8        | -1.3k        | -3.42       | 0.55        | -2.87       | <u>5.49</u> | 2.07        | -517.4       | -2.38        | -1.85       | -4.23       | <b>5.88</b> | <b>3.5</b>  |
| ANCHOR                                                      | <b>9.5</b>   | <b>0.33</b> | -2.61        | <b>-2.61</b> | <u>4.69</u> | <b>4.69</b> | <b>1k</b>    | <b>2.75</b> | <b>8.96</b> | <b>8.96</b> | <b>5.73</b> | <b>5.73</b> | <b>189.1</b> | <b>0.87</b>  | <u>3.28</u> | <b>3.28</b> | 3.13        | <u>3.13</u> |
| Harmless category                                           |              |             |              |              |             |             |              |             |             |             |             |             |              |              |             |             |             |             |
| <i>Llama-3.2-1B-Instruct</i> → <i>Llama-3.1-8B-Instruct</i> |              |             |              |              |             |             |              |             |             |             |             |             |              |              |             |             |             |             |
| Method                                                      | AHar         |             |              |              |             |             | PKU          |             |             |             |             |             | RAIL         |              |             |             |             |             |
|                                                             | In-dist.     |             | PKU          |              | RAIL        |             | In-dist.     |             | AHar        |             | RAIL        |             | In-dist.     |              | AHar        |             | PKU         |             |
|                                                             | PGR (%)      | WRG         | AOG          | NTS          | AOG         | NTS         | PGR (%)      | WRG         | AOG         | NTS         | AOG         | NTS         | PGR (%)      | WRG          | AOG         | NTS         | AOG         | NTS         |
| Naive                                                       | 122.6        | 3.09        | 1.97         | 1.97         | <u>0</u>    | <u>0</u>    | 146.1        | 4.91        | <u>2.65</u> | <u>2.65</u> | 2.26        | 2.26        | <u>55.7</u>  | <u>1.03</u>  | <u>2.44</u> | <u>2.44</u> | <u>2.04</u> | <u>2.04</u> |
| Conf.-based                                                 | 129.4        | 3.26        | <u>3.11</u>  | <b>3.11</b>  | -0.31       | -0.31       | 151.2        | 5.08        | 2.61        | 2.61        | 2.47        | 2.47        | -72.4        | -1.34        | 0.79        | -0.55       | 0.32        | -1.02       |
| SEAM                                                        | -264.3       | -6.66       | <b>4.5</b>   | -2.16        | -0.51       | -7.17       | <b>236.3</b> | <b>7.94</b> | 1.64        | 1.64        | <b>6.27</b> | <b>6.27</b> | -277.8       | -5.14        | -4.74       | -9.88       | -0.25       | -5.39       |
| ANCHOR                                                      | <b>134.5</b> | <b>3.39</b> | 2.38         | <u>2.38</u>  | <b>0.31</b> | <b>0.31</b> | <u>163.4</u> | <u>5.49</u> | <b>3.43</b> | <b>3.43</b> | <u>4.62</u> | <u>4.62</u> | <b>77.8</b>  | <b>1.44</b>  | <b>2.62</b> | <b>2.62</b> | <b>2.37</b> | <b>2.37</b> |
| <i>Qwen3-1.7B</i> → <i>Qwen3-8B</i>                         |              |             |              |              |             |             |              |             |             |             |             |             |              |              |             |             |             |             |
| Method                                                      | AHar         |             |              |              |             |             | PKU          |             |             |             |             |             | RAIL         |              |             |             |             |             |
|                                                             | In-dist.     |             | PKU          |              | RAIL        |             | In-dist.     |             | AHar        |             | RAIL        |             | In-dist.     |              | AHar        |             | PKU         |             |
|                                                             | PGR (%)      | WRG         | AOG          | NTS          | AOG         | NTS         | PGR (%)      | WRG         | AOG         | NTS         | AOG         | NTS         | PGR (%)      | WRG          | AOG         | NTS         | AOG         | NTS         |
| Naive                                                       | <b>160.9</b> | <b>1.48</b> | <u>2.05</u>  | <b>2.05</b>  | <u>0.52</u> | <u>0.52</u> | 1k           | 0.82        | <u>1.35</u> | <u>1.35</u> | <b>1.85</b> | <b>1.85</b> | <u>-19.1</u> | <u>-0.51</u> | <u>1.53</u> | 1.02        | 1.72        | 1.21        |
| Conf.-based                                                 | 5.4          | 0.05        | -1.72        | -1.72        | -0.51       | -0.51       | 1.5k         | 1.23        | -0.39       | -0.39       | <u>1.54</u> | <u>1.54</u> | -30.7        | -0.82        | <b>1.88</b> | <u>1.06</u> | <b>3.93</b> | <b>3.11</b> |
| SEAM                                                        | -1.6k        | -14.71      | <b>5.4</b>   | -9.31        | -1.54       | -16.25      | <b>4.2k</b>  | <b>3.36</b> | -8.49       | -8.49       | -2.67       | -2.67       | -385         | -10.28       | -10.79      | -21.07      | 1.07        | -9.21       |
| ANCHOR                                                      | <b>160.9</b> | <b>1.48</b> | 0.66         | <u>0.66</u>  | <b>1.65</b> | <b>1.65</b> | <u>2.7k</u>  | <u>2.21</u> | <b>1.43</b> | <b>1.43</b> | <b>1.85</b> | <b>1.85</b> | <b>0</b>     | <b>0</b>     | 1.4         | <b>1.4</b>  | <u>2.46</u> | <u>2.46</u> |

H3 = HelpSteer3, AH = Anthropic Helpful, UF = UltraFeedback, AHar = Anthropic Harmless, In-dist. = In-distribution metrics. The best value in each column and model block is shown in bold, and the second-best value is underlined. PGR is reported for comparability with prior W2S work. Because it is a ratio-based metric, it can take very large negative values when a method performs far below the weak supervisor or when the weak-to-strong gap is small.

anchoring strengths and layer choices, suggesting that representation preservation alone cannot eliminate all source-target preference mismatch.

A comparison with the strong-supervised ceiling further contextualizes these remaining failures. The mean ID/OOD ceiling gaps are 1.01/6.37 for Naive, 1.56/6.27 for confidence-based W2S, 8.73/8.05 for SEAM, and only -0.11/5.45 for ANCHOR. Moreover, the additional OOD gap attributable to weak supervision is only 0.23 points for ANCHOR, compared with 1.15, 1.05, and 2.83 for the three base-lines. This suggests that much of the remaining OOD deficit reflects source-target transfer difficulty that persists even under gold supervision.

Table 2: Comparison with preservation and optimization controls on Llama/HelpSteer3. Full hyperparameter sweeps are reported in the appendix.

| Method                 | WRG         | AH          |             | UF          |             |
|------------------------|-------------|-------------|-------------|-------------|-------------|
|                        |             | AOG         | NTS         | AOG         | NTS         |
| Naive                  | 1.71        | 1.46        | 1.46        | 4.80        | 4.80        |
| WI ( $\alpha = 0.75$ ) | 1.38        | 4.12        | 4.12        | 5.32        | 5.32        |
| Early stopping         | 2.47        | 3.80        | 3.80        | 6.26        | 6.26        |
| LoRA ( $r = 8$ )       | 1.49        | 2.86        | 2.86        | 6.14        | 6.14        |
| L2-SP                  | 2.69        | 4.16        | 4.16        | <b>9.03</b> | <b>9.03</b> |
| ANCHOR                 | <b>3.57</b> | <b>8.19</b> | <b>8.19</b> | 7.93        | 7.93        |

**Comparison with preservation controls.** To test whether ANCHOR’s gains can be reproduced by simply restricting fine-tuning, we com-

pare against weight interpolation (WI), source-validation early stopping, reduced-rank LoRA, and L2-SP parameter-space regularization in the same Llama/HelpSteer3 setting (Table 2). The simpler controls provide partial improvements over Naive W2S, but none matches ANCHOR’s overall ID–OOD balance. In particular, early stopping and rank-8 LoRA reach UltraFeedback AOGs of 6.26 and 6.14, respectively, but remain substantially below ANCHOR on Anthropic Helpful. Parameter-space preservation is more competitive: L2-SP with  $\lambda = 10^{-4}$  reaches 9.03 AOG on UltraFeedback, exceeding ANCHOR’s 7.93, but obtains lower WRG (2.69 vs. 3.57) and substantially weaker transfer to Anthropic Helpful (4.16 vs. 8.19 AOG). Together, these results suggest that preserving pretrained information is beneficial, while directly constraining hidden representations provides a stronger overall preservation transfer balance than reducing optimization, limiting adaptation capacity, or constraining parameter displacement alone. Full WI and L2-SP sweeps are reported in Appendix F.

## 6 Ablation Studies

We conduct ablation studies to better understand the role of representation anchoring in ANCHOR. Unless otherwise stated, all ablations are conducted with the Llama model family using HelpSteer3 as the source training domain. Specifically, we analyze the effect of the anchoring strength and the choice of anchored layer.

### 6.1 Effect of Anchoring Strength

We first study the sensitivity of ANCHOR to the anchoring coefficient  $\lambda$ . We vary  $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}\}$ , which controls the strength of the representation-anchoring loss relative to the weak preference objective. A larger  $\lambda$  creates a stronger constraint on the student representations, which may better preserve pretrained knowledge but can also weaken adaptation to the source preference signal.

Figure 3 shows that decreasing the anchoring coefficient consistently improves performance across WRG, AOG, and NTS. The consistent trend suggests that overly strong anchoring can constrain adaptation, whereas a weaker anchoring penalty better balances weak preference learning and representation preservation.

Table 3: Effect of anchored layer with  $\lambda = 10^{-4}$ . Bold indicates the best value in each column.

| Method       | WRG         | Anthropic Helpful |             | UltraFeedback |             |
|--------------|-------------|-------------------|-------------|---------------|-------------|
|              |             | AOG               | NTS         | AOG           | NTS         |
| Middle-layer | 2.14        | 4.42              | 4.42        | <b>8.22</b>   | <b>8.22</b> |
| ANCHOR       | <b>3.57</b> | <b>8.19</b>       | <b>8.19</b> | 7.93          | 7.93        |

### 6.2 Effect of Anchored Layer

We further study the choice of anchoring layer. Final-layer states are closer to the reward head and may be more strongly shaped by source-domain preference learning, whereas middle-layer states may preserve more general transferable features while leaving higher layers free to adapt.

Table 3 shows that ANCHOR provides a better overall balance across in-distribution and transfer settings. While middle-layer anchoring is slightly stronger on UltraFeedback, ANCHOR performs better in-distribution and transfers more strongly to Anthropic Helpful, suggesting that the chosen anchored layer better preserves transferable representations without overly restricting adaptation.

## 7 Related Work

**Weak-to-strong generalization and scalable oversight.** Recent work has applied weak-to-strong (W2S) generalization to alignment and preference learning, showing that weak aligned models or weak supervisors can improve stronger models through preference optimization and related alignment objectives (Zhu et al., 2025; Lyu et al., 2025). At the same time, these studies also expose limitations of weak supervision, including overfitting to weak signals or inheriting supervisor errors (Shi et al., 2025; Yang et al., 2025b). Beyond preference supervision, W2S has also been studied in reasoning, mathematical problem solving, weak-guided search, and agentic decision-making, suggesting that weak supervision is a broader paradigm for eliciting strong-model capabilities (Yang et al., 2024; Zhou et al., 2024; Nie et al., 2025).

### Preference learning and reward modeling.

Learning scalar reward models from comparisons is a core component of RLHF-style alignment (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). Most reward-modeling objectives use a Bradley–Terry likelihood over response pairs (Bradley and Terry, 1952), and recent benchmarks further treat reward models as

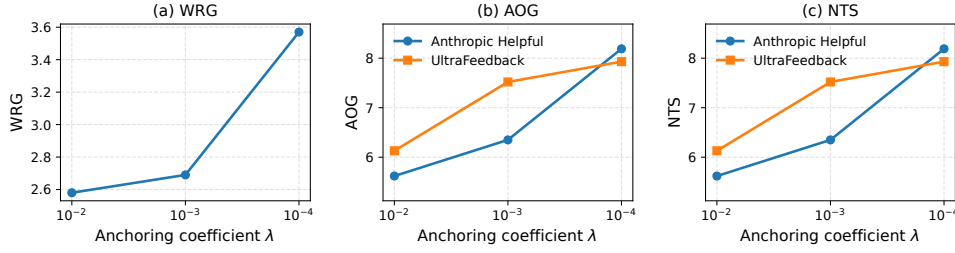


Figure 3: Effect of anchoring strength  $\lambda$  on in-distribution and out-of-distribution transfer performance. Larger x-axis values correspond to smaller anchoring coefficients, i.e., weaker representation anchoring.

standalone evaluators across chat, reasoning, and safety settings (Lambert et al., 2025). In parallel, direct preference-optimization methods bypass explicit reward-model training for policy optimization (Rafailov et al., 2023; Gheshlaghi Azar et al., 2024; Hong et al., 2024; Meng et al., 2024). Joint evaluator models instead compare candidate responses together, including pairwise rankers, comparative LLM judges, and evaluator LMs with pairwise-ranking modes (Jiang et al., 2023; Liusie et al., 2024; Li et al., 2024b; Kim et al., 2024); however, pair-dependent scoring is less compatible with settings that require independent candidate scores.

**Representation-preserving fine-tuning.** Fine-tuning can degrade pretrained features and OOD robustness (Kumar et al., 2022; Li et al., 2024c; Zang et al., 2024), related to catastrophic forgetting in continual learning (Li et al., 2024a; Zheng et al., 2025). Prior approaches preserve knowledge through distillation (Hinton et al., 2015), continual-learning regularization (Kirkpatrick et al., 2017; Li and Hoiem, 2017; Wu et al., 2024), or parameter-space constraints (Lee et al., 2020). Unlike continual learning, which typically preserves performance on previously observed tasks, our setting evaluates zero-shot transfer from a single source preference domain to unseen domains. ANCHOR adapts representation preservation to W2S reward modeling by anchoring response-token hidden states to a frozen pretrained reference, with no inference-time modification.

## 8 Conclusion

We study weak-to-strong reward modeling under zero-shot preference-domain shift and introduce an evaluation protocol that trains on one preference domain while testing both in-distribution and on unseen target domains. Using this protocol, we show that in-distribution W2S gains do not necessarily indicate robust alignment: meth-

ods that improve source-domain performance may still transfer poorly across domains. To address this gap, we propose ANCHOR, a representation-anchoring method regularizing the strong student toward a frozen pretrained reference during weak-supervised training. Across preference domains and model families, ANCHOR better preserves in-distribution performance while improving out-of-distribution transfer. Our findings suggest that robust W2S alignment should be evaluated not only by whether strong models learn from weak supervision, but also by whether the learned reward function remains reliable beyond the training domain.

## Limitations

Our study focuses primarily on offline pairwise reward-model accuracy across a fixed suite of helpful and harmless preference datasets. While this controlled setting allows us to systematically analyze preference-domain transfer, offline metrics may not fully capture downstream deployment behavior. We provide initial downstream validation through best-of-4 reranking (Appendix G), where ANCHOR improves candidate selection across both ID and OOD settings. However, broader downstream evaluation through online RLHF policy optimization, DPO-style policy training, and open-ended generation is left for future work.

Our cross-dataset evaluation also represents a compound preference-domain shift: the datasets may differ simultaneously in prompt distributions, response styles and sources, annotation protocols, and preference policies. Consequently, our experiments do not disentangle the contribution of each factor to OOD failure, nor establish representation drift as its sole cause. Isolating these sources of shift through more controlled evaluations is left for future work.

Furthermore, we note a distinct trade-off regarding computational efficiency. In a representative Llama/UltraFeedback run, ANCHOR increases

training time and GPU-hours by approximately 14% relative to Naive W2S (4.11 vs. 3.62 minutes; 0.069 vs. 0.060 GPU-hours), while measured peak VRAM remains effectively unchanged (38.66 vs. 38.71 GB/GPU). This overhead comes from one additional frozen-reference forward pass, with no reference-model backward pass and no inference-time overhead. Finally, our empirical validation is limited to specific helpful/harmless domains and specific model scales; future work is required to evaluate ANCHOR across broader preference axes, more diverse task domains, and larger model families.

## Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-PhD-2025-08-058).

## References

- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. [Concrete problems in AI safety](#). *Preprint*, arXiv:1606.06565.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *Preprint*, arXiv:2204.05862.
- Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiuūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, and 27 others. 2022. [Measuring progress on scalable oversight for large language models](#). *Preprint*, arXiv:2211.03540.
- Ralph Allan Bradley and Milton E. Terry. 1952. [Rank analysis of incomplete block designs: The method of paired comparisons](#). *Biometrika*, 39(3/4):324–345.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. [Weak-to-strong generalization: Eliciting strong capabilities with weak supervision](#). *Preprint*, arXiv:2312.09390.
- Moses Charikar, Chirag Pabbaraju, and Kirankumar Shrivastava. 2024. [Quantifying the gain in weak-to-strong generalization](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. [Deep reinforcement learning from human preferences](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. [ULTRA FEEDBACK: Boosting language models with scaled AI feedback](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9722–9744. PMLR.
- Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. [A general theoretical paradigm to understand learning from human preferences](#). In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4447–4455. PMLR.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#). *Preprint*, arXiv:1503.02531.
- Jiwoo Hong, Noah Lee, and James Thorne. 2024. [ORPO: Monolithic preference optimization without reference model](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.
- Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. 2025. [PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31983–32016, Vienna, Austria. Association for Computational Linguistics.
- Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. [LLM-blender: Ensembling large language models with pairwise ranking and generative fusion](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. [Prometheus 2: An open source language model specialized in evaluating other language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharmashan Kumaran, and Raia Hadsell. 2017. [Overcoming catastrophic forgetting in neural networks](#). *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.
- Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2022. [Fine-tuning can distort pretrained features and underperform out-of-distribution](#). In *International Conference on Learning Representations*.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. [RewardBench: Evaluating reward models for language modeling](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.
- Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. 2020. [Mixout: Effective regularization to finetune large-scale pretrained language models](#). In *International Conference on Learning Representations*.
- Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. [Scalable agent alignment via reward modeling: a research direction](#). Preprint, arXiv:1811.07871.
- Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024a. [Revisiting catastrophic forgetting in large language model tuning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.
- Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2024b. [Generative judge for evaluating alignment](#). In *The Twelfth International Conference on Learning Representations*.
- Kaican Li, Weiyan Xie, Yongxiang Huang, Didan Deng, Lanqing Hong, Zhenguo Li, Ricardo Silva, and Nevin L. Zhang. 2024c. [Dual risk minimization: Towards next-level robustness in fine-tuning zero-shot models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 66025–66057. Curran Associates, Inc.
- Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024d. [Superfiltering: Weak-to-strong data filtering for fast instruction-tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14255–14273, Bangkok, Thailand. Association for Computational Linguistics.
- Yongqi Li, Xin Miao, Mayi Xu, and Tiejun Qian. 2025. [Strong empowered and aligned weak mastered annotation for weak-to-strong generalization](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27437–27445.
- Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.
- Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. [LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian’s, Malta. Association for Computational Linguistics.
- Yongang Lyu, Lingyong Yan, Zihan Wang, Dawei Yin, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2025. [MACPO: Weak-to-strong alignment via multi-agent contrastive preference optimization](#). In *The Thirteenth International Conference on Learning Representations*.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. [SimPO: Simple preference optimization with a reference-free reward](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Fan Nie, Lan Feng, Haotian Ye, Weixin Liang, Pan Lu, Huaxiu Yao, Alexandre Alahi, and James Zou. 2025. [Weak-for-strong: Training weak meta-agent to harness strong executors](#). In *Second Conference on Language Modeling*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Mishra, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). Preprint, arXiv:2203.02155.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in neural information processing systems*, volume 36, pages 53728–53741.
- Junhao Shi, Qinyuan Cheng, Zhaoye Fei, Yining Zheng, Qipeng Guo, and Xipeng Qiu. 2025. [How to mitigate overfitting in weak-to-strong generalization?](#) In

- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16100–16118, Vienna, Austria. Association for Computational Linguistics.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in neural information processing systems*, volume 33, pages 3008–3021.
- Sumit Verma, Pritam Prasun, Arpit Jaiswal, and Pritish Kumar. 2025. [RAIL in the wild: Operationalizing responsible AI evaluation using Anthropic’s value dataset](#). *Preprint*, arXiv:2505.00204.
- Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. 2025. [HelpSteer3-Preference: Open human-annotated preference data across diverse tasks and languages](#). *Preprint*, arXiv:2505.11475.
- Yichen Wu, Long-Kai Huang, Renzhen Wang, Deyu Meng, and Ying Wei. 2024. [Meta continual learning revisited: Implicitly enhancing online Hessian approximation via variance reduction](#). In *The Twelfth International Conference on Learning Representations*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Wenkai Yang, Shiqi Shen, Guangyao Shen, Wei Yao, Yong Liu, Gong Zhi, Yankai Lin, and Ji-Rong Wen. 2025b. [Super\(ficial\)-alignment: Strong models may deceive weak models in weak-to-strong generalization](#). In *The Thirteenth International Conference on Learning Representations*.
- Yuqing Yang, Yan Ma, and Pengfei Liu. 2024. [Weak-to-strong reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8350–8367, Miami, Florida, USA. Association for Computational Linguistics.
- Yuhang Zang, Hanlin Goh, Joshua M. Susskind, and Chen Huang. 2024. [Overcoming the pitfalls of vision-language model finetuning for OOD generalization](#). In *The Twelfth International Conference on Learning Representations*.
- Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. 2025. [Spurious forgetting in continual learning of language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Zhanhui Zhou, Zhixuan Liu, Jie Liu, Zhichen Dong, Chao Yang, and Yu Qiao. 2024. [Weak-to-strong search: Align large language models via searching over small language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Wenhong Zhu, Zhiwei He, Xiaofeng Wang, Pengfei Liu, and Rui Wang. 2025. [Weak-to-strong preference optimization: Stealing reward from weak aligned model](#). In *The Thirteenth International Conference on Learning Representations*.

## A Hyperparameters and Implementation Details

We use LoRA fine-tuning for all weak-to-strong methods, including both the weak teacher and the strong student models. Unless otherwise specified, we use the final checkpoint after three epochs for evaluation and apply the same training configuration across methods:

- LoRA rank: 16
- Learning rate:  $2 \times 10^{-4}$
- Number of epochs: 3
- Anchoring coefficient for ANCHOR:  $\lambda = 10^{-4}$

For ANCHOR, we select  $\lambda = 10^{-4}$  from a preliminary sweep over  $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}\}$  on the HelpSteer3 source-domain setting, where it yields the strongest in-domain validation performance. We then fix this value across all datasets and model families to avoid per-setting hyperparameter tuning.

All experiments were conducted on NVIDIA RTX A6000 GPUs. Reproducing the reported results requires approximately 2400 GPU-hours in total, including runs across datasets, random seeds, and baselines. To quantify the additional cost of representation anchoring, we report training efficiency for a representative Llama/UltraFeedback run in Table 4.

Table 4: Training overhead on a representative Llama/UltraFeedback run. Relative time and VRAM are normalized to Naive W2S. ANCHOR adds approximately 14% training time and GPU-hours, while measured peak VRAM remains effectively unchanged.

| Method      | Time (min) | VRAM (GB) | GPU-hours | Relative Time | Relative VRAM |
|-------------|------------|-----------|-----------|---------------|---------------|
| Naive W2S   | 3.62       | 38.71     | 0.060     | 1.00×         | 1.00×         |
| Conf.-based | 3.48       | 38.71     | 0.058     | 0.96×         | 1.00×         |
| ANCHOR      | 4.11       | 38.66     | 0.069     | 1.14×         | 1.00×         |

## B Data Processing Pipeline

Most preference-learning datasets provide a prompt, a chosen response, a rejected response, and the corresponding chosen and rejected scores. We first remove samples where the chosen and rejected scores are equal, since these tied examples do not provide a clear preference signal. We then concatenate each prompt with its chosen and rejected responses to construct a chosen conversation and

a rejected conversation, respectively. These paired conversations are used as the input examples for reward model training and evaluation.

## C Dataset Statistics and Preference-Domain Shift Diagnostics

### C.1 Dataset split statistics

Table 5 summarizes the dataset statistics used in our experiments, including the train, validation, test, weak-teacher training, and weak-to-strong training splits for each preference dataset.

Table 5: Dataset statistics used in our weak-to-strong reward-modeling experiments. The training split is divided equally into weak-teacher training data and weak-to-strong training data.

| Dataset            | Category | Train   | Validation | Test  | Weak-teacher train | W2S train |
|--------------------|----------|---------|------------|-------|--------------------|-----------|
| Anthropic Helpful  | Helpful  | 115,396 | 11,540     | 2,332 | 57,698             | 57,698    |
| HelpSteer3         | Helpful  | 17,708  | 1,771      | 915   | 8,854              | 8,854     |
| UltraFeedback      | Helpful  | 53,748  | 5,375      | 1,728 | 26,874             | 26,874    |
| Anthropic Harmless | Harmless | 42,254  | 4,226      | 2,298 | 21,127             | 21,127    |
| PKU-SafeRLHF       | Harmless | 26,874  | 2,688      | 1,222 | 13,437             | 13,437    |
| RAIL               | Harmless | 7,862   | 887        | 973   | 3,931              | 3,931     |

### C.2 Cross-dataset shift diagnostics

To characterize the distribution differences underlying our cross-dataset evaluation, Table 6 reports simple statistics of the prompt and preferred-response distributions. Token lengths are measured using the Llama-3 tokenizer. Even within the same broad preference category, these datasets differ substantially in prompt and response length. For example, HelpSteer3 contains considerably longer prompts and preferred responses than Anthropic Helpful, while the harmlessness datasets also differ in both response length and the frequency with which the preferred response is longer than the rejected response.

Table 6: Cross-dataset preference-domain diagnostics. Prompt and chosen-response lengths are mean token counts measured with the Llama-3 tokenizer. “Chosen longer” denotes the fraction of preference pairs in which the preferred response is longer than the rejected response.

| Dataset            | Prompt len. | Chosen len. | Chosen longer |
|--------------------|-------------|-------------|---------------|
| Anthropic Helpful  | 162         | 56          | 59%           |
| HelpSteer3         | 565         | 389         | 56%           |
| UltraFeedback      | 182         | 275         | 56%           |
| Anthropic Harmless | 149         | 38          | 43%           |
| PKU-SafeRLHF       | 61          | 82          | 38%           |
| RAIL               | 137         | 47          | 34%           |

The datasets additionally differ in response provenance and annotation policy. Anthropic HH-

RLHF is based on binary human preference comparisons. Meanwhile, HelpSteer3 uses graded preferences collected from multiple annotators and UltraFeedback derives preferences from multidimensional GPT-4-based evaluations. In the harmlessness category, PKU-SafeRLHF focuses on preferences between safer and less-safe responses while RAIL uses multidimensional responsible-AI criteria. Consequently, our cross-dataset setting jointly varies prompt distributions, response styles and sources, and preference-labeling policies.

These diagnostics support treating the evaluation as a compound preference-domain shift rather than an ordinary held-out dataset split. However, because these factors vary jointly, they do not identify any single factor, including representation drift, response-distribution shift, or annotation-policy differences, as the sole cause of the observed OOD degradation.

## D Full Accuracy Results and Strong-Ceiling Analysis

### D.1 Full Accuracy Results

Table 7 reports the full accuracy results with standard deviations across different reward-modeling methods, datasets, and model families.

### D.2 Gap to Strong-Supervised Ceiling

To further contextualize weak-to-strong performance, we compare each student with a strong reward model trained directly on gold labels from the evaluation domain. For a student trained on source domain  $S$  and evaluated on domain  $T$ , we define the target-ceiling gap as

$$G_{S \rightarrow T} = A(m_T^{\text{gt}}, T) - A(m_S^{\text{w2s}}, T),$$

where  $m_T^{\text{gt}}$  denotes the strong model trained directly on gold labels from  $T$ . When  $T = S$ , this reduces to the in-distribution strong-ceiling gap; when  $T \neq S$ , it measures how far a zero-shot source-trained student remains from direct target-domain supervision.

Table 8 shows that ANCHOR has the smallest mean gap to the strong-supervised reference under both ID and OOD evaluation. Its mean ID gap is  $-0.11$  points, slightly exceeding the source-domain strong-supervised reference on average, while its mean OOD target-ceiling gap is  $5.45$  points, compared with  $6.37$ ,  $6.27$ , and  $8.05$  for

Naive W2S, confidence-based W2S, and SEAM, respectively. ANCHOR also obtains the lowest gap in  $9.5/12$  ID and  $11/24$  OOD cases under fractional tie counting.

To distinguish limitations due to cross-domain transfer from those due specifically to weak supervision, we further decompose the OOD gap as

$$G_{S \rightarrow T} = \underbrace{A(m_T^{\text{gt}}, T) - A(m_S^{\text{gt}}, T)}_{\text{source-target gold-transfer gap}} + \underbrace{A(m_S^{\text{gt}}, T) - A(m_S^{\text{w2s}}, T)}_{\text{weak-supervision gap}}.$$

The first term measures the difficulty of transferring even gold source-domain supervision to the unseen target domain, whereas the second measures the additional gap introduced by weak-supervised training when the source dataset and strong-model architecture are held fixed. Averaged across OOD evaluations, the source-target gold-transfer gap is  $5.22$  points. The additional weak-supervision gap is only  $0.23$  points for ANCHOR, compared with  $1.15$ ,  $1.05$ , and  $2.83$  for Naive W2S, confidence-based W2S, and SEAM, respectively. ANCHOR substantially reduces the remaining discrepancy associated with weak-label training, while much of its residual OOD gap reflects source-target transfer difficulty that persists even under gold source-domain supervision.

## E Anchoring Ablations and Representation Analysis

### E.1 Expanded $\lambda$ and layer ablations

The main-text ablations in Section 6 use Llama with HelpSteer3 as the source domain. Here, we extend both the anchoring-strength and layer-selection analyses to two additional Qwen settings: HelpSteer3 and UltraFeedback as source domains. These experiments test whether the conclusions from the main ablation persist across model families and source datasets.

**Anchoring strength.** Table 9 reports the raw pairwise accuracies for  $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}\}$ . In addition to the Llama/HelpSteer3 setting reported in the main paper, we evaluate Qwen trained on UltraFeedback and HelpSteer3. Importantly,  $\lambda = 10^{-4}$  was selected using the original Llama/HelpSteer3 experiment and then fixed for the Qwen experiments rather than tuned independently for each source-target pair.

Across these settings,  $\lambda = 10^{-4}$  obtains the highest or tied-highest accuracy in all nine ID/OOD comparisons. Stronger anchoring generally reduces performance, suggesting that excessive preservation can restrict adaptation to weak preference supervision. At the same time, the difficult Qwen HelpSteer3→Anthropic Helpful transfer remains weak even at  $\lambda = 10^{-4}$ , indicating that this failure cannot be explained solely by overly strong anchoring.

**Anchoring layer.** We next compare final-layer ANCHOR with middle-layer anchoring at the fixed coefficient  $\lambda = 10^{-4}$ . Table 10 extends the main-text layer ablation to the same two Qwen settings.

Final-layer anchoring achieves higher accuracy in eight of the nine comparisons, although several differences are small. Middle-layer anchoring is slightly better for Llama HelpSteer3→UltraFeedback, while final-layer anchoring provides the stronger overall ID/OOD balance across the evaluated model families and source domains. These results support our use of final-layer ANCHOR as the default variant, while also indicating that the optimal anchoring location can vary for individual source-target pairs.

## E.2 Representation Drift Analysis

We further analyze how weak-supervised fine-tuning changes the internal representations of the strong model. Specifically, we compute CKA and CCA distances between each fine-tuned LoRA checkpoint and the frozen Llama-8B base model at different layers. Lower distance indicates that the fine-tuned model preserves representations closer to the pretrained base model, while higher distance indicates larger representation drift.

Figure 4 shows that representation drift is concentrated in middle and final layers, which is expected because these layers are closer to the reward head and are more directly affected by preference learning. Compared with naive and confidence-based weak-to-strong training, ANCHOR generally reduces CKA/CCA distance to the pretrained base model, indicating that anchoring constrains excessive representational movement during weak-label training.

Importantly, the best-performing anchored checkpoint is not the one with the smallest representation distance. Very strong anchoring, such as  $\lambda = 10^{-2}$ , keeps the representations closer to the base model, but can overly restrict adaptation

to the source preference task. In contrast, the best-performing checkpoint,  $\lambda = 10^{-4}$ , allows moderate representation movement while still reducing drift relative to unanchored baselines. This supports a controlled-drift interpretation: ANCHOR improves weak-to-strong reward modeling not by freezing pretrained representations, but by balancing representation preservation with task-relevant adaptation.

## E.3 Preference-Relevant Linear Probing

The CKA/CCA analysis measures geometric representation drift, but greater similarity to the pretrained representation space does not necessarily imply preservation of information relevant to preference prediction. We therefore conduct a complementary linear-probe analysis. We freeze the representations of Llama reward models trained on HelpSteer3 using Naive W2S, confidence-based W2S, and ANCHOR, and train the same linear probe using gold HelpSteer3 preference labels. We then evaluate the probe zero-shot on gold test labels from HelpSteer3, Anthropic Helpful, and UltraFeedback, without target-domain adaptation.

As shown in Table 11, ANCHOR achieves the highest probe accuracy on both the source dataset and the two unseen target datasets. In particular, it improves over Naive W2S by 4.12 points on HelpSteer3, 4.89 points on Anthropic Helpful, and 0.92 points on UltraFeedback. These results indicate that the representations preserved by ANCHOR contain more linearly accessible information predictive of gold preferences, complementing the geometric evidence from CKA/CCA. However, the probe captures aggregate preference-relevant information reflected in pairwise labels; it does not separately identify individual factors such as refusal behavior, verbosity, or other preference attributes.

## F Preservation and Optimization Controls

### F.1 Representation-Space versus Parameter-Space Preservation

We additionally compare ANCHOR with L2-SP, a parameter-space regularization baseline that penalizes deviation from the pretrained initialization. The comparison is conducted when training on HelpSteer3 and evaluating zero-shot transfer to Anthropic Helpful and UltraFeedback, as reported in Table 12. The full sweep shows that L2-SP can improve transfer in some cases, espe-

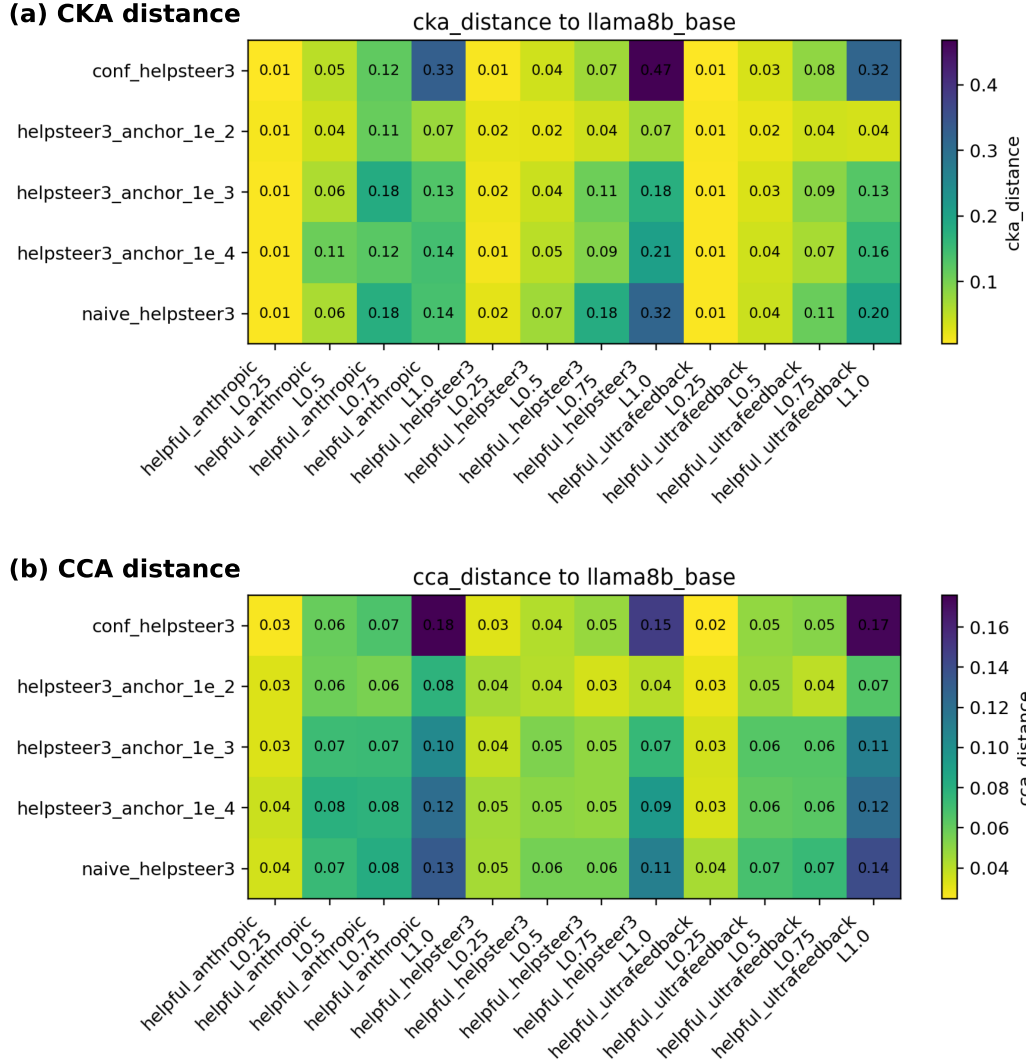


Figure 4: CKA and CCA distances between HelpSteer3-trained LoRA checkpoints and the frozen Llama-8B base model across evaluation domains and layers.

cially on UltraFeedback, confirming that preserving pretrained information is useful. However, its effect is less balanced across target domains: the strongest L2-SP setting on UltraFeedback substantially underperforms ANCHOR on Anthropic Helpful and yields lower WRG. In contrast, ANCHOR achieves the best WRG and the strongest average transfer across the two target domains, suggesting that representation-space preservation provides a more stable mechanism than parameter-space regularization alone.

## F.2 Weight Interpolation and Optimization Controls

We further test whether ANCHOR’s gains can be reproduced by simply limiting the magnitude, duration, or capacity of fine-tuning. We consider

three controls: (1) weight interpolation (WI), which scales the Naive W2S LoRA update toward the pretrained initialization; (2) source-validation-based early stopping; and (3) reduced-capacity LoRA with rank  $r = 8$ , compared with the default rank  $r = 16$ . We evaluate these controls on both Llama and Qwen with HelpSteer3 as the source domain, and additionally evaluate WI on Llama/RAIL to cover the harmless category.

Across these settings, the simpler controls do not reproduce ANCHOR’s overall ID/OOD balance, as depicted in Table 13. Weight interpolation provides partial improvements for Llama/HelpSteer3 at  $\alpha = 0.75$ , but smaller interpolation coefficients substantially degrade both source and target performance; the same pattern is more pronounced for Qwen/HelpSteer3 and Llama/RAIL. Source-

validation early stopping and rank-8 LoRA also remain below ANCHOR in the HelpSteer3 experiments. In the difficult Qwen/HelpSteer3 setting, neither control restores positive source-domain gain, whereas ANCHOR achieves positive WRG and stronger transfer to UltraFeedback. Together with the L2-SP comparison above, these results indicate that ANCHOR’s gains are not explained solely by smaller parameter updates, shorter optimization, or reduced adaptation capacity.

## **G Downstream Best-of-4 Reranking Evaluation**

To examine whether improvements in pairwise reward-model accuracy translate to downstream candidate selection, we conduct a best-of-4 reranking evaluation on UltraFeedback using the Llama reward models. We evaluate models trained on HelpSteer3, Anthropic Helpful, and UltraFeedback. The first two settings therefore measure zero-shot OOD transfer to UltraFeedback, while training on UltraFeedback provides the corresponding ID setting. We compare Naive W2S, confidence-based W2S, and ANCHOR, and report both best-of-4 top-1 accuracy and pairwise accuracy in Table 14.

ANCHOR achieves the highest best-of-4 top-1 accuracy in all three settings, including both OOD transfers. Averaged across source domains, it improves top-1 accuracy by 2.22 points over Naive W2S and 1.81 points over confidence-based W2S. The same ordering is observed for pairwise accuracy. These results provide initial evidence that the transfer improvements observed in offline reward-model evaluation can translate into better downstream candidate selection. Broader evaluation through RLHF, DPO-style policy training, and open-ended generation is left for future work.

Table 7: Accuracy results across helpful and harmless preference-domain shifts. Each three-column block indicates the source training dataset, and each subcolumn indicates the evaluation dataset. Standard deviations over three seeds are shown as subscripts.

| Helpful                                                     |                         |                       |                       |                        |                       |                       |                       |                       |                       |
|-------------------------------------------------------------|-------------------------|-----------------------|-----------------------|------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| <i>Llama-3.2-1B-Instruct</i> → <i>Llama-3.1-8B-Instruct</i> |                         |                       |                       |                        |                       |                       |                       |                       |                       |
| Method                                                      | Src: HelpSteer3         |                       |                       | Src: Anthropic Helpful |                       |                       | Src: UltraFeedback    |                       |                       |
|                                                             | HelpSteer3              | AHelpful              | UltraFB               | AHelpful               | HelpSteer3            | UltraFB               | UltraFB               | HelpSteer3            | AHelpful              |
| Weak teacher                                                | 74.03 <sub>0.77</sub>   | 52.96 <sub>0.59</sub> | 63.89 <sub>0.23</sub> | 69.47 <sub>0.51</sub>  | 64.26 <sub>0.23</sub> | 65.51 <sub>0.72</sub> | 74.94 <sub>0.49</sub> | 64.48 <sub>0.72</sub> | 62.14 <sub>0.72</sub> |
| Strong ceiling                                              | 79.34 <sub>0.37</sub>   | 60.98 <sub>0.77</sub> | 73.50 <sub>0.40</sub> | 72.64 <sub>0.59</sub>  | 67.76 <sub>0.72</sub> | 71.47 <sub>0.66</sub> | 77.31 <sub>0.26</sub> | 71.26 <sub>0.64</sub> | 64.67 <sub>0.77</sub> |
| Naive                                                       | 75.74 <sub>0.72</sub>   | 54.42 <sub>0.40</sub> | 68.69 <sub>0.71</sub> | 72.60 <sub>0.69</sub>  | 68.85 <sub>0.72</sub> | 70.49 <sub>0.77</sub> | 76.68 <sub>0.28</sub> | 70.27 <sub>0.31</sub> | 62.74 <sub>0.77</sub> |
| Confidence-based                                            | 71.91 <sub>0.72</sub>   | 61.02 <sub>0.24</sub> | 71.53 <sub>0.45</sub> | 71.70 <sub>0.19</sub>  | 68.09 <sub>0.77</sub> | 70.25 <sub>0.72</sub> | 75.93 <sub>0.21</sub> | 67.87 <sub>0.72</sub> | 62.39 <sub>0.72</sub> |
| SEAM                                                        | 58.36 <sub>0.77</sub>   | 62.91 <sub>0.70</sub> | 70.14 <sub>0.50</sub> | 56.43 <sub>0.77</sub>  | 58.25 <sub>0.77</sub> | 65.05 <sub>0.59</sub> | 68.29 <sub>0.76</sub> | 58.80 <sub>0.72</sub> | 62.31 <sub>0.75</sub> |
| ANCHOR                                                      | 77.60 <sub>0.38</sub>   | 61.15 <sub>0.52</sub> | 71.82 <sub>0.43</sub> | 72.86 <sub>0.47</sub>  | 66.34 <sub>0.77</sub> | 71.12 <sub>0.72</sub> | 77.84 <sub>0.29</sub> | 70.49 <sub>0.58</sub> | 66.55 <sub>0.72</sub> |
| <i>Qwen3-1.7B</i> → <i>Qwen3-8B</i>                         |                         |                       |                       |                        |                       |                       |                       |                       |                       |
| Method                                                      | Src: HelpSteer3         |                       |                       | Src: Anthropic Helpful |                       |                       | Src: UltraFeedback    |                       |                       |
|                                                             | HelpSteer3              | AHelpful              | UltraFB               | AHelpful               | HelpSteer3            | UltraFB               | UltraFB               | HelpSteer3            | AHelpful              |
| Weak teacher                                                | 75.74 <sub>0.72</sub>   | 59.43 <sub>0.32</sub> | 67.42 <sub>0.41</sub> | 71.05 <sub>0.65</sub>  | 61.64 <sub>0.16</sub> | 68.58 <sub>0.72</sub> | 76.97 <sub>0.26</sub> | 67.21 <sub>0.50</sub> | 62.09 <sub>0.72</sub> |
| Strong ceiling                                              | 79.23 <sub>0.72</sub>   | 58.40 <sub>0.72</sub> | 71.88 <sub>0.25</sub> | 71.31 <sub>0.72</sub>  | 65.36 <sub>0.25</sub> | 75.23 <sub>0.53</sub> | 77.43 <sub>0.62</sub> | 69.73 <sub>0.77</sub> | 62.99 <sub>0.34</sub> |
| Naive                                                       | 72.68 <sub>0.72</sub>   | 59.39 <sub>0.27</sub> | 67.59 <sub>0.72</sub> | 71.50 <sub>0.16</sub>  | 67.21 <sub>0.35</sub> | 73.73 <sub>0.77</sub> | 76.56 <sub>0.15</sub> | 70.82 <sub>0.50</sub> | 63.46 <sub>0.18</sub> |
| Confidence-based                                            | 73.44 <sub>0.49</sub>   | 57.85 <sub>0.35</sub> | 71.82 <sub>0.77</sub> | 72.77 <sub>0.72</sub>  | 68.42 <sub>0.49</sub> | 73.50 <sub>0.55</sub> | 76.79 <sub>0.33</sub> | 68.63 <sub>0.72</sub> | 65.35 <sub>0.20</sub> |
| SEAM                                                        | 64.15 <sub>0.22</sub>   | 67.63 <sub>0.77</sub> | 73.21 <sub>0.49</sub> | 67.63 <sub>0.72</sub>  | 62.19 <sub>0.72</sub> | 74.07 <sub>0.72</sub> | 74.59 <sub>0.19</sub> | 65.36 <sub>0.20</sub> | 67.97 <sub>0.33</sub> |
| ANCHOR                                                      | 76.07 <sub>0.72</sub>   | 56.82 <sub>0.77</sub> | 72.11 <sub>0.77</sub> | 73.80 <sub>0.76</sub>  | 70.60 <sub>0.72</sub> | 74.31 <sub>0.56</sub> | 77.84 <sub>0.72</sub> | 69.73 <sub>0.72</sub> | 65.22 <sub>0.33</sub> |
| Harmless                                                    |                         |                       |                       |                        |                       |                       |                       |                       |                       |
| <i>Llama-3.2-1B-Instruct</i> → <i>Llama-3.1-8B-Instruct</i> |                         |                       |                       |                        |                       |                       |                       |                       |                       |
| Method                                                      | Src: Anthropic Harmless |                       |                       | Src: PKU               |                       |                       | Src: RAIL             |                       |                       |
|                                                             | AHarmless               | PKU                   | RAIL                  | PKU                    | AHarmless             | RAIL                  | RAIL                  | AHarmless             | PKU                   |
| Weak teacher                                                | 70.02 <sub>0.77</sub>   | 90.75 <sub>0.72</sub> | 88.49 <sub>0.37</sub> | 85.92 <sub>0.64</sub>  | 61.58 <sub>0.46</sub> | 76.98 <sub>0.56</sub> | 89.21 <sub>0.62</sub> | 68.62 <sub>0.21</sub> | 93.62 <sub>0.50</sub> |
| Strong ceiling                                              | 72.54 <sub>0.65</sub>   | 92.64 <sub>0.46</sub> | 90.24 <sub>0.33</sub> | 89.28 <sub>0.77</sub>  | 64.88 <sub>0.55</sub> | 81.19 <sub>0.72</sub> | 91.06 <sub>0.47</sub> | 70.23 <sub>0.69</sub> | 96.48 <sub>0.72</sub> |
| Naive                                                       | 73.11 <sub>0.39</sub>   | 92.72 <sub>0.72</sub> | 88.49 <sub>0.43</sub> | 90.83 <sub>0.77</sub>  | 64.23 <sub>0.72</sub> | 79.24 <sub>0.72</sub> | 90.24 <sub>0.77</sub> | 71.06 <sub>0.33</sub> | 95.66 <sub>0.45</sub> |
| Confidence-based                                            | 73.28 <sub>0.72</sub>   | 93.86 <sub>0.65</sub> | 88.18 <sub>0.62</sub> | 91.00 <sub>0.35</sub>  | 64.19 <sub>0.50</sub> | 79.45 <sub>0.70</sub> | 87.87 <sub>0.77</sub> | 69.41 <sub>0.35</sub> | 93.94 <sub>0.72</sub> |
| SEAM                                                        | 63.36 <sub>0.69</sub>   | 95.25 <sub>0.68</sub> | 87.98 <sub>0.54</sub> | 93.86 <sub>0.50</sub>  | 63.22 <sub>0.77</sub> | 83.25 <sub>0.72</sub> | 84.07 <sub>0.55</sub> | 63.88 <sub>0.72</sub> | 93.37 <sub>0.33</sub> |
| ANCHOR                                                      | 73.41 <sub>0.57</sub>   | 93.37 <sub>0.21</sub> | 89.00 <sub>0.35</sub> | 91.41 <sub>0.44</sub>  | 65.01 <sub>0.72</sub> | 81.60 <sub>0.22</sub> | 90.65 <sub>0.24</sub> | 71.24 <sub>0.54</sub> | 95.34 <sub>0.77</sub> |
| <i>Qwen3-1.7B</i> → <i>Qwen3-8B</i>                         |                         |                       |                       |                        |                       |                       |                       |                       |                       |
| Method                                                      | Src: Anthropic Harmless |                       |                       | Src: PKU               |                       |                       | Src: RAIL             |                       |                       |
|                                                             | AHarmless               | PKU                   | RAIL                  | PKU                    | AHarmless             | RAIL                  | RAIL                  | AHarmless             | PKU                   |
| Weak teacher                                                | 71.80 <sub>0.72</sub>   | 87.97 <sub>0.28</sub> | 82.73 <sub>0.77</sub> | 87.97 <sub>0.48</sub>  | 63.58 <sub>0.77</sub> | 81.50 <sub>0.27</sub> | 89.31 <sub>0.41</sub> | 67.75 <sub>0.56</sub> | 91.24 <sub>0.72</sub> |
| Strong ceiling                                              | 72.72 <sub>0.33</sub>   | 93.21 <sub>0.67</sub> | 87.46 <sub>0.53</sub> | 88.05 <sub>0.42</sub>  | 64.53 <sub>0.72</sub> | 81.50 <sub>0.77</sub> | 91.98 <sub>0.72</sub> | 69.58 <sub>0.58</sub> | 95.34 <sub>0.77</sub> |
| Naive                                                       | 73.28 <sub>0.24</sub>   | 90.02 <sub>0.77</sub> | 83.25 <sub>0.34</sub> | 88.79 <sub>0.46</sub>  | 64.93 <sub>0.46</sub> | 83.35 <sub>0.72</sub> | 88.80 <sub>0.72</sub> | 69.28 <sub>0.72</sub> | 92.96 <sub>0.35</sub> |
| Confidence-based                                            | 71.85 <sub>0.21</sub>   | 86.25 <sub>0.45</sub> | 82.22 <sub>0.19</sub> | 89.20 <sub>0.77</sub>  | 63.19 <sub>0.77</sub> | 83.04 <sub>0.51</sub> | 88.49 <sub>0.72</sub> | 69.63 <sub>0.30</sub> | 95.17 <sub>0.47</sub> |
| SEAM                                                        | 57.09 <sub>0.31</sub>   | 93.37 <sub>0.72</sub> | 81.19 <sub>0.20</sub> | 91.33 <sub>0.77</sub>  | 55.09 <sub>0.72</sub> | 78.83 <sub>0.70</sub> | 79.03 <sub>0.76</sub> | 56.96 <sub>0.19</sub> | 92.31 <sub>0.61</sub> |
| ANCHOR                                                      | 73.28 <sub>0.77</sub>   | 88.46 <sub>0.53</sub> | 84.38 <sub>0.72</sub> | 90.18 <sub>0.32</sub>  | 65.01 <sub>0.67</sub> | 82.43 <sub>0.34</sub> | 89.31 <sub>0.77</sub> | 69.15 <sub>0.71</sub> | 93.70 <sub>0.72</sub> |

AHelpful = Anthropic Helpful, UltraFB = UltraFeedback, AHarmless = Anthropic Harmless

Table 8: Gap between weak-to-strong students and strong models trained directly on gold labels from the evaluation domain. ID and OOD columns report mean target-ceiling gaps in accuracy points. “Weak-sup. gap” denotes the additional OOD gap attributable to replacing gold source supervision with weak supervision. Lowest-gap counts use fractional counting for ties.

| Method      | ID gap ↓     | OOD gap ↓   | Weak-sup. gap ↓ | Lowest ID       | Lowest OOD     |
|-------------|--------------|-------------|-----------------|-----------------|----------------|
| Naive W2S   | 1.01         | 6.37        | 1.15            | 0.5 / 12        | 4 / 24         |
| Conf.-based | 1.56         | 6.27        | 1.05            | 0 / 12          | 2 / 24         |
| SEAM        | 8.73         | 8.05        | 2.83            | 2 / 12          | 7 / 24         |
| ANCHOR      | <b>-0.11</b> | <b>5.45</b> | <b>0.23</b>     | <b>9.5 / 12</b> | <b>11 / 24</b> |

Table 9: Expanded ablation on the anchoring coefficient  $\lambda$ . We report pairwise accuracy (%) in the evaluation order shown in the second column. H3, AH, and UF denote HelpSteer3, Anthropic Helpful, and UltraFeed-back, respectively.

| Model / source | Eval. order  | $\lambda = 10^{-2}$   | $\lambda = 10^{-3}$   | $\lambda = 10^{-4}$          |
|----------------|--------------|-----------------------|-----------------------|------------------------------|
| Llama / H3     | H3 / AH / UF | 76.61 / 58.21 / 70.22 | 76.72 / 60.06 / 70.58 | <b>77.60 / 61.15 / 71.82</b> |
| Qwen / UF      | UF / AH / H3 | 77.37 / 60.55 / 69.73 | 77.84 / 62.95 / 70.16 | <b>77.84 / 65.22 / 70.49</b> |
| Qwen / H3      | H3 / AH / UF | 73.33 / 54.31 / 71.31 | 73.33 / 54.93 / 71.64 | <b>76.07 / 56.82 / 72.11</b> |

Table 10: Expanded comparison of middle- and final-layer anchoring at  $\lambda = 10^{-4}$ . Values are pairwise accuracy (%).

| Model / source | Eval. order  | Middle layer                 | Final layer                  |
|----------------|--------------|------------------------------|------------------------------|
| Llama / H3     | H3 / AH / UF | 76.17 / 57.38 / <b>72.11</b> | <b>77.60 / 61.15 / 71.82</b> |
| Qwen / UF      | UF / AH / H3 | 77.64 / 65.05 / 69.01        | <b>77.84 / 65.22 / 70.49</b> |
| Qwen / H3      | H3 / AH / UF | 75.05 / 55.43 / 71.73        | <b>76.07 / 56.82 / 72.11</b> |

Table 11: Linear-probe accuracy (%) on frozen representations of HelpSteer3-trained reward models. The probe is trained using gold HelpSteer3 preference labels and evaluated zero-shot on the three helpfulness datasets.

| Method          | HelpSteer3   | Anthropic Helpful | UltraFeedback |
|-----------------|--------------|-------------------|---------------|
| Naive W2S       | 71.07        | 55.06             | 69.33         |
| Conf.-based W2S | 70.19        | 58.96             | 70.05         |
| ANCHOR          | <b>75.19</b> | <b>59.95</b>      | <b>70.25</b>  |

Table 12: Comparison between representation-space anchoring and parameter-space preservation when training on HelpSteer3.

| Method               | WRG         | Anthropic Helpful |             | UltraFeedback |             |
|----------------------|-------------|-------------------|-------------|---------------|-------------|
|                      |             | AOG               | NTS         | AOG           | NTS         |
| ANCHOR ( $10^{-2}$ ) | 2.58        | 5.25              | 5.25        | 6.33          | 6.33        |
| ANCHOR ( $10^{-3}$ ) | 2.69        | 7.10              | 7.10        | 6.69          | 6.69        |
| ANCHOR ( $10^{-4}$ ) | <b>3.57</b> | <b>8.19</b>       | <b>8.19</b> | <b>7.93</b>   | <b>7.93</b> |
| L2-SP ( $10^{-3}$ )  | 2.30        | 5.63              | 5.63        | 7.33          | 7.33        |
| L2-SP ( $10^{-4}$ )  | 2.69        | 4.16              | 4.16        | <b>9.03</b>   | <b>9.03</b> |
| L2-SP ( $10^{-5}$ )  | 2.26        | 6.16              | 6.16        | 7.58          | 7.58        |

Table 13: Comparison with weight interpolation and optimization/capacity controls. For HelpSteer3, Target 1 and Target 2 are Anthropic Helpful and UltraFeed-back; for RAIL, they are Anthropic Harmless and PKU-SafeRLHF, respectively.

| Model / source     | Method                        | WRG         | Target 1     |              | Target 2    |             |
|--------------------|-------------------------------|-------------|--------------|--------------|-------------|-------------|
|                    |                               |             | AOG          | NTS          | AOG         | NTS         |
| Llama / HelpSteer3 | WI ( $\alpha = 0.25$ )        | -22.01      | -0.30        | -22.31       | -15.39      | -37.40      |
|                    | WI ( $\alpha = 0.50$ )        | -17.20      | 3.73         | -13.47       | -10.01      | -27.21      |
|                    | WI ( $\alpha = 0.75$ )        | 1.38        | 4.12         | 4.12         | 5.32        | 5.32        |
|                    | WI ( $\alpha = 1.00$ ; Naive) | 1.71        | 1.46         | 1.46         | 4.80        | 4.80        |
|                    | Early stopping                | 2.47        | 3.80         | 3.80         | 6.26        | 6.26        |
|                    | LoRA ( $r = 8$ )              | 1.49        | 2.86         | 2.86         | 6.14        | 6.14        |
|                    | ANCHOR                        | <b>3.57</b> | <b>8.19</b>  | <b>8.19</b>  | <b>7.93</b> | <b>7.93</b> |
| Qwen / HelpSteer3  | WI ( $\alpha = 0.25$ )        | -26.01      | -5.74        | -31.75       | -16.49      | -42.50      |
|                    | WI ( $\alpha = 0.50$ )        | -21.20      | -4.51        | -25.71       | -8.51       | -29.71      |
|                    | WI ( $\alpha = 0.75$ )        | -6.23       | -4.00        | -10.23       | -0.12       | -6.35       |
|                    | WI ( $\alpha = 1.00$ ; Naive) | -3.06       | -0.04        | -3.10        | 0.17        | -2.89       |
|                    | Early stopping                | -3.84       | -3.17        | -7.01        | -0.58       | -4.42       |
|                    | LoRA ( $r = 8$ )              | -3.73       | -3.72        | -7.45        | -3.29       | -7.02       |
|                    | ANCHOR                        | <b>0.33</b> | <b>-2.61</b> | <b>-2.61</b> | <b>4.69</b> | <b>4.69</b> |
| Llama / RAIL       | WI ( $\alpha = 0.25$ )        | -31.35      | -13.62       | -44.97       | -33.15      | -64.50      |
|                    | WI ( $\alpha = 0.50$ )        | -4.22       | -2.48        | -6.70        | -1.89       | -6.11       |
|                    | WI ( $\alpha = 0.75$ )        | 0.43        | 1.94         | 1.94         | 0.80        | 0.80        |
|                    | WI ( $\alpha = 1.00$ ; Naive) | 1.03        | 2.44         | 2.44         | 2.04        | 2.04        |
|                    | ANCHOR                        | <b>1.44</b> | <b>2.62</b>  | <b>2.62</b>  | <b>2.37</b> | <b>2.37</b> |

Table 14: Best-of-4 reranking on UltraFeedback using Llama reward models. HelpSteer3 and Anthropic Helpful are OOD source domains, while UltraFeedback is the ID source. Higher is better.

| Source            | Method      | Bo4 Top-1    | Pairwise     |
|-------------------|-------------|--------------|--------------|
| HelpSteer3        | Naive       | 45.17        | 69.58        |
|                   | Conf.-based | 47.09        | 70.53        |
|                   | ANCHOR      | <b>48.86</b> | <b>72.14</b> |
| Anthropic Helpful | Naive       | 44.18        | 70.31        |
|                   | Conf.-based | 44.21        | 70.51        |
|                   | ANCHOR      | <b>45.24</b> | <b>70.68</b> |
| UltraFeedback     | Naive       | 53.55        | 76.37        |
|                   | Conf.-based | 52.84        | 75.45        |
|                   | ANCHOR      | <b>55.47</b> | <b>77.44</b> |