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

Anonymous ACL submission

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 for a representational failure mode in which weak-supervised fine-tuning can pull 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 consistently 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.

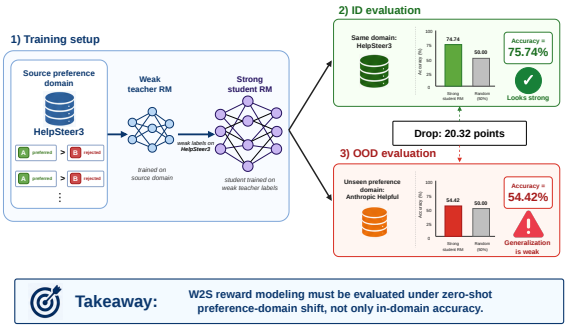

Figure 1: **ID success can hide OOD failure.** Although the resulting student can perform well on in-domain evaluation, its performance can degrade substantially when tested on unseen preference domains. This motivates evaluating W2S reward models under zero-shot preference-domain shift rather than relying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

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). 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. 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 compo-

nent 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 address this fragility, we propose *Representation Anchoring* (ANCHOR), a simple yet effective regularization method that 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 yields the most consistent gains 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^{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

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

and the standard reward-modeling objective is

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

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_{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^{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

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

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_θ using only k_i^{train}. We evaluate r_θ in-distribution on the held-out test set from the same domain,

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

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

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

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

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

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

where $\mathrm{Acc}(\cdot, E_{\mathrm{test}})$ 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:

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

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:

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

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

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

and define the Net Transfer Score (NTS) as

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

When the strong student preserves or improves source-domain performance relative to the weak supervisor, $C_{\mathrm{ID}}(S; m) = 0$, 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 ($\mathcal{L}_{anchor}$), allowing the student to learn from weak supervision while discouraging source-domain fine-tuning from distorting representations that are useful for 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 directly observe the ground-truth preference label. Instead, a weak teacher reward model w_S, trained on the source domain S, provides a soft preference label

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

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

The strong student m_S predicts

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

and is trained to match the weak teacher through the soft preference loss

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

This objective trains the strong student to match the weak teacher's soft preference distribution while maintaining the standard scalar reward modeling form: each prompt–response pair receives an independent score $m_S(x, y)$, and pairwise preferences are derived from score differences. In ANCHOR, this loss provides the preference-learning signal and is combined with the representation-anchoring regularizer to improve zero-shot transfer across preference domains.

4.2 Representation Anchoring Regularizer

Our central hypothesis is that the pretrained strong model already contains broadly useful representations for preference judgment, but fine-tuning on weak labels from a single source domain can pull these representations toward source-specific artifacts. 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_{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

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

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

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

to exclude prompt tokens and padding tokens from the anchoring loss. Following that, the masked hidden-state distance at layer ℓ is defined as

$$D_{\mathrm{anchor}}^\ell(x, y) = \frac{\sum_{t=1}^T \mathrm{m}_t(x, y)\left\|\Delta_t^\ell(x, y)\right\|_2^2}{d\sum_{t=1}^T \mathrm{m}_t(x, y)},$$

where

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

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

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

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

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

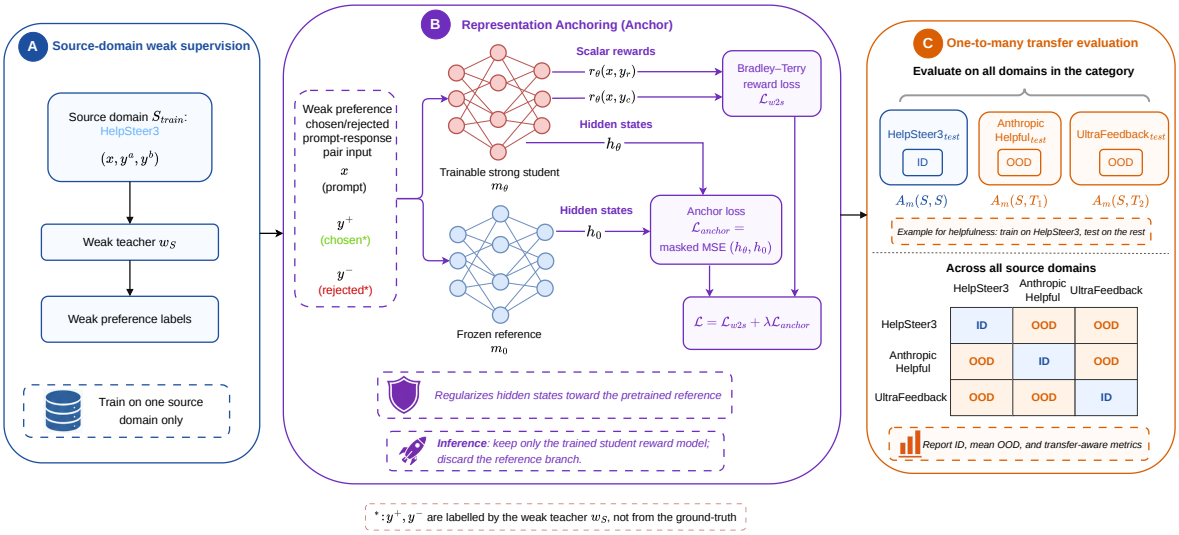

Figure 2: **Overview of ANCHOR under zero-shot preference-domain shift.** ANCHOR trains a strong reward model with the standard weak-to-strong preference loss $\mathcal{L}_{\text{w2s}}$, while regularizing its response-token hidden states toward a frozen pretrained reference model through $\mathcal{L}_{\text{anchor}}$. 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.

where λ 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 preference-relevant 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:

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

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

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

The corresponding training objective now becomes

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

Middle-Layer ANCHOR is useful both as an alternative regularizer and as a diagnostic. If preserving intermediate representations improves OOD transfer, it supports the view that weak-to-strong failures under preference-domain shift are not explained only by noisy weak labels, but also by representation drift during source-domain fine-tuning. 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. 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 the 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. These results suggest that representation anchoring provides a more robust signal for weak-to-strong transfer under preference-domain shift.

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 λ. 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 λ 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 co-

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

Llama-3.2-1B-Instruct → Llama-3.1-8B-Instruct

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	32.23	1.71	1.46	1.46	4.8	4.8	98.7	3.13	**4.59**	**4.59**	4.98	4.98	73.4	1.74	5.79	5.79	0.6	0.6
Conf.-based	-39.9	-2.12	8.06	5.94	7.64	5.52	70.3	2.23	3.83	3.83	4.74	4.74	41.8	0.99	3.39	3.39	0.25	0.25
SEAM	-295.1	-15.67	**9.95**	-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	**67.2**	**3.57**	8.19	**8.19**	**7.93**	**7.93**	**106.9**	**3.39**	2.08	2.08	**5.61**	**5.61**	**97.9**	**2.32**	**6.01**	**6.01**	**4.41**	**4.41**

Qwen3-1.7B → Qwen3-8B

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	-0.04	-3.1	0.17	-2.89	173.1	0.45	5.57	5.57	5.15	5.15	-89.1	-0.41	**3.61**	3.2	1.37	0.96
Conf.-based	-65.9	-2.3	-1.58	-3.88	4.4	2.1	661.5	1.72	6.78	6.78	4.92	4.92	-39.1	-0.18	1.42	1.24	3.26	3.08
SEAM	-332.1	-11.59	**8.2**	-3.39	**5.79**	-5.8	$-1.3k$	-3.42	0.55	-2.87	5.49	2.07	-517.4	-2.38	-1.85	-4.23	**5.88**	**3.5**
ANCHOR	**9.5**	**0.33**	-2.61	**-2.61**	4.69	**4.69**	**1k**	**2.75**	**8.96**	**8.96**	**5.73**	**5.73**	**189.1**	**0.87**	3.28	**3.28**	3.13	3.13

Harmless category

Llama-3.2-1B-Instruct → Llama-3.1-8B-Instruct

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	0	0	146.1	4.91	2.65	2.65	2.26	2.26	55.7	1.03	2.44	2.44	2.04	2.04
Conf.-based	129.4	3.26	3.11	**3.11**	-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	**4.5**	-2.16	-0.51	-7.17	**236.3**	**7.94**	1.64	1.64	**6.27**	**6.27**	-277.8	-5.14	-4.74	-9.88	-0.25	-5.39
ANCHOR	**134.5**	**3.39**	2.38	2.38	**0.31**	**0.31**	163.4	5.49	**3.43**	**3.43**	4.62	4.62	**77.8**	**1.44**	**2.62**	**2.62**	**2.37**	**2.37**

Qwen3-1.7B → Qwen3-8B

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	**160.9**	**1.48**	2.05	**2.05**	0.52	0.52	$1k$	0.82	1.35	1.35	**1.85**	**1.85**	-19.1	-0.51	1.53	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	1.54	1.54	-30.7	-0.82	**1.88**	1.06	**3.93**	**3.11**
SEAM	-1.6k	-14.71	**5.4**	-9.31	-1.54	-16.25	**4.2k**	**3.36**	-8.49	-8.49	-2.67	-2.67	-385	-10.28	-10.79	-21.07	1.07	-9.21
ANCHOR	**160.9**	**1.48**	0.66	0.66	**1.65**	**1.65**	2.7k	2.21	**1.43**	**1.43**	**1.85**	**1.85**	**0**	**0**	1.4	**1.4**	2.46	2.46

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.

efficient 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.

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.

Method	WRG	Anthropic Helpful		UltraFeedback	
		AOG	NTS	AOG	NTS
MLA	2.14	4.42	4.42	**8.22**	**8.22**
ANCHOR	**3.57**	**8.19**	**8.19**	7.93	7.93

Table 2: Effect of anchored layer with $\lambda = 10^{-4}$. Bold indicates the best value in each column.

Table 2 shows that ANCHOR provides a better overall balance across in-distribution and transfer settings. While MLA is slightly stronger on UltraFeedback, ANCHOR performs better in-distribution and transfers more strongly to Anthropic Helpful, suggesting that the chosen anchored layer bet-

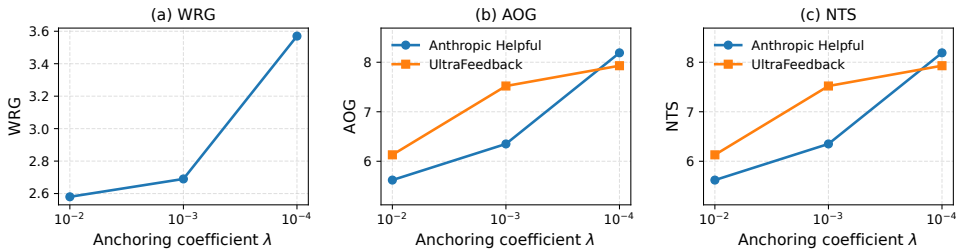

Figure 3: Effect of anchoring strength λ on in-distribution and out-of-distribution transfer performance. Larger x-axis values correspond to smaller anchoring coefficients, i.e., weaker representation anchoring.

ter 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 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 improve source-domain accuracy while degrading pretrained features and OOD robustness (Kumar et al., 2022; Li et al., 2024c; Zang et al., 2024). Similar concerns arise in LLM tuning, where adaptation can cause catastrophic or spurious forgetting of prior capabilities (Li et al., 2024a; Zheng et al., 2025). Existing remedies preserve information through distillation (Hinton et al., 2015), continual-learning regularization (Kirkpatrick et al., 2017; Li and Hoiem, 2017), or parameter-space constraints toward pretrained weights (Lee et al., 2020). ANCHOR adapts this idea to W2S reward modeling by anchoring hidden states to a frozen pretrained reference, reducing representation drift while keeping the scalar reward-model interface unchanged at inference.

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: methods 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 on offline pairwise reward-model accuracy across a fixed suite of helpful and harmless preference datasets. While this tightly controlled setting is essential for cleanly isolating and analyzing preference-domain transfer, we recognize that offline proxy metrics may not fully capture the nuances of downstream deployment. Specifically, our evaluation does not account for best-of-n selection mechanics, online RLHF policy optimization, or the final quality of open-ended text generation.

Furthermore, we note a distinct trade-off regarding computational efficiency. While ANCHOR requires no architectural modifications or extra overhead at inference time, thus maintaining the standard reward model interface, it does introduce an additional reference model forward pass during training. This increases the VRAM and time requirements for optimization. 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.

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 Lukošiūtė, 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 Shiragur. 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. ULTRAFEEDBACK: 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, Dharshan 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 Tieyun 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.

Yougang 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 Slama, 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.

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. For all training runs, we use the final checkpoint after three epochs for evaluation. Unless otherwise specified, we use the same training configuration across methods:

- LoRA rank: 16
- Learning rate: 2×10^{-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. The runtime varies with dataset size and method implementation.

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

Table 3 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.

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

Table 3: 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.

D Full Accuracy Results

Table 4 reports the full accuracy results with standard deviations across different reward-modeling methods, datasets, and model families.

E 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 slightly drift off 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.

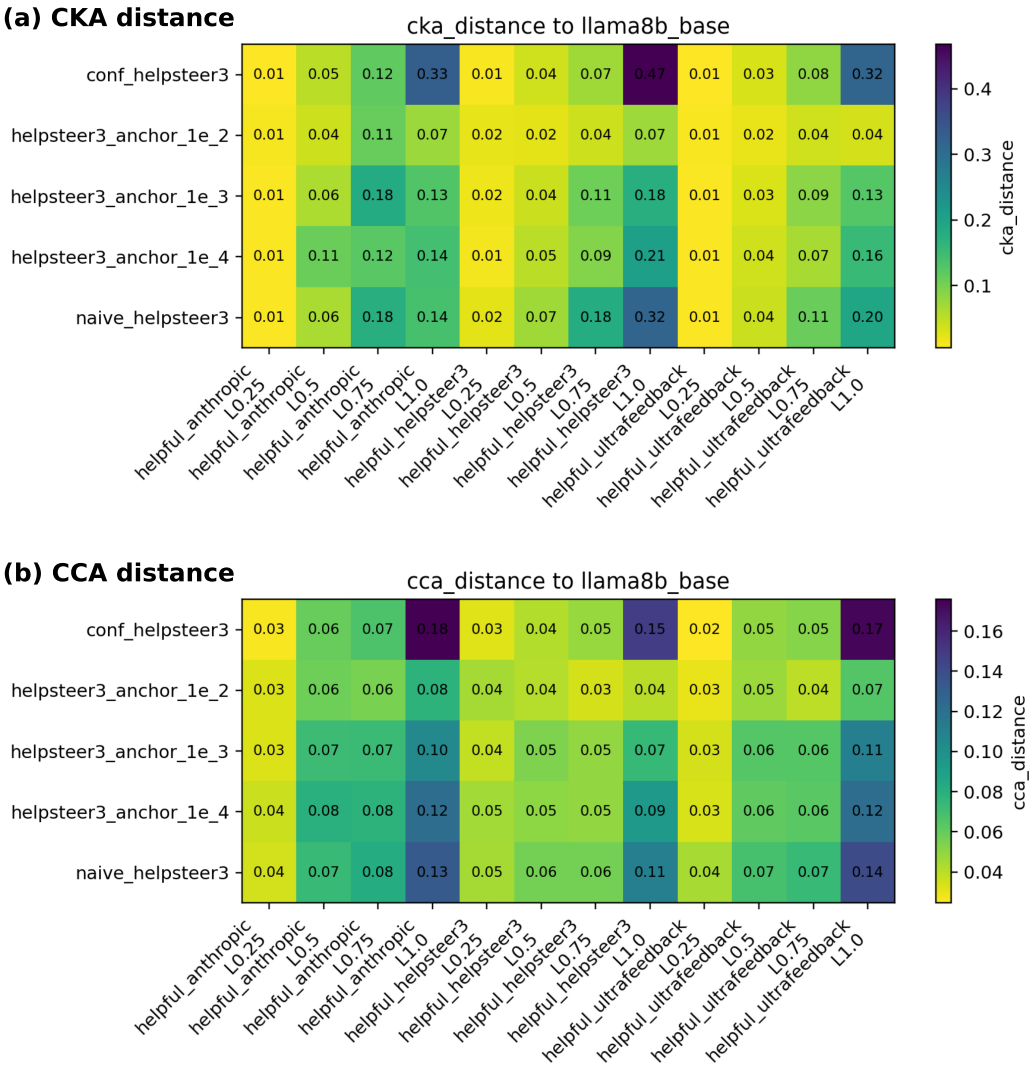

Figure 4: CKA and CCA distances between HelpSteer3-trained LoRA checkpoints and the frozen Llama-8B base model across evaluation domains and layers.

F Representation-Space versus Parameter-Space Preservation

Comparison with 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 5. The full sweep shows that L2-SP can improve transfer in some cases, especially 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.

Table 4: 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

Llama-3.2-1B-Instruct → Llama-3.1-8B-Instruct

	Src: HelpSteer3			**Src: Anthropic Helpful**			**Src: UltraFeedback**		
Method	HelpSteer3	AHelpful	UltraFB	AHelpful	HelpSteer3	UltraFB	UltraFB	HelpSteer3	AHelpful
Weak teacher	$74.03_{0.77}$	$52.96_{0.59}$	$63.89_{0.23}$	$69.47_{0.51}$	$64.26_{0.23}$	$65.51_{0.72}$	$74.94_{0.49}$	$64.48_{0.72}$	$62.14_{0.72}$
Strong ceiling	$79.34_{0.37}$	$60.98_{0.77}$	$73.50_{0.40}$	$72.64_{0.59}$	$67.76_{0.72}$	$71.47_{0.66}$	$77.31_{0.26}$	$71.26_{0.64}$	$64.67_{0.77}$
Naive	$75.74_{0.72}$	$54.42_{0.40}$	$68.69_{0.71}$	$72.60_{0.69}$	$68.85_{0.72}$	$70.49_{0.77}$	$76.68_{0.28}$	$70.27_{0.31}$	$62.74_{0.77}$
Confidence-based	$71.91_{0.72}$	$61.02_{0.24}$	$71.53_{0.45}$	$71.70_{0.19}$	$68.09_{0.77}$	$70.25_{0.72}$	$75.93_{0.21}$	$67.87_{0.72}$	$62.39_{0.72}$
SEAM	$58.36_{0.77}$	$62.91_{0.70}$	$70.14_{0.50}$	$56.43_{0.77}$	$58.25_{0.77}$	$65.05_{0.59}$	$68.29_{0.76}$	$58.80_{0.72}$	$62.31_{0.75}$
ANCHOR	$77.60_{0.38}$	$61.15_{0.52}$	$71.82_{0.43}$	$72.86_{0.47}$	$66.34_{0.77}$	$71.12_{0.72}$	$77.84_{0.29}$	$70.49_{0.58}$	$66.55_{0.72}$

Qwen3-1.7B → Qwen3-8B

	Src: HelpSteer3			**Src: Anthropic Helpful**			**Src: UltraFeedback**		
Method	HelpSteer3	AHelpful	UltraFB	AHelpful	HelpSteer3	UltraFB	UltraFB	HelpSteer3	AHelpful
Weak teacher	$75.74_{0.72}$	$59.43_{0.32}$	$67.42_{0.41}$	$71.05_{0.65}$	$61.64_{0.16}$	$68.58_{0.72}$	$76.97_{0.26}$	$67.21_{0.50}$	$62.09_{0.72}$
Strong ceiling	$79.23_{0.72}$	$58.40_{0.72}$	$71.88_{0.25}$	$71.31_{0.72}$	$65.36_{0.25}$	$75.23_{0.53}$	$77.43_{0.62}$	$69.73_{0.77}$	$62.99_{0.34}$
Naive	$72.68_{0.72}$	$59.39_{0.27}$	$67.59_{0.72}$	$71.50_{0.16}$	$67.21_{0.35}$	$73.73_{0.77}$	$76.56_{0.15}$	$70.82_{0.50}$	$63.46_{0.18}$
Confidence-based	$73.44_{0.49}$	$57.85_{0.35}$	$71.82_{0.77}$	$72.77_{0.72}$	$68.42_{0.49}$	$73.50_{0.55}$	$76.79_{0.33}$	$68.63_{0.72}$	$65.35_{0.20}$
SEAM	$64.15_{0.22}$	$67.63_{0.77}$	$73.21_{0.49}$	$67.63_{0.72}$	$62.19_{0.72}$	$74.07_{0.72}$	$74.59_{0.19}$	$65.36_{0.20}$	$67.97_{0.33}$
ANCHOR	$76.07_{0.72}$	$56.82_{0.77}$	$72.11_{0.77}$	$73.80_{0.76}$	$70.60_{0.72}$	$74.31_{0.56}$	$77.84_{0.72}$	$69.73_{0.72}$	$65.22_{0.33}$

Harmless

Llama-3.2-1B-Instruct → Llama-3.1-8B-Instruct

	Src: Anthropic Harmless			**Src: PKU**			**Src: RAIL**		
Method	AHarmless	PKU	RAIL	PKU	AHarmless	RAIL	RAIL	AHarmless	PKU
Weak teacher	$70.02_{0.77}$	$90.75_{0.72}$	$88.49_{0.37}$	$85.92_{0.64}$	$61.58_{0.46}$	$76.98_{0.56}$	$89.21_{0.62}$	$68.62_{0.21}$	$93.62_{0.50}$
Strong ceiling	$72.54_{0.65}$	$92.64_{0.46}$	$90.24_{0.33}$	$89.28_{0.77}$	$64.88_{0.55}$	$81.19_{0.72}$	$91.06_{0.47}$	$70.23_{0.69}$	$96.48_{0.72}$
Naive	$73.11_{0.39}$	$92.72_{0.72}$	$88.49_{0.43}$	$90.83_{0.77}$	$64.23_{0.72}$	$79.24_{0.72}$	$90.24_{0.77}$	$71.06_{0.33}$	$95.66_{0.45}$
Confidence-based	$73.28_{0.72}$	$93.86_{0.65}$	$88.18_{0.62}$	$91.00_{0.35}$	$64.19_{0.50}$	$79.45_{0.70}$	$87.87_{0.77}$	$69.41_{0.35}$	$93.94_{0.72}$
SEAM	$63.36_{0.69}$	$95.25_{0.68}$	$87.98_{0.54}$	$93.86_{0.50}$	$63.22_{0.77}$	$83.25_{0.72}$	$84.07_{0.55}$	$63.88_{0.72}$	$93.37_{0.33}$
ANCHOR	$73.41_{0.57}$	$93.37_{0.21}$	$89.00_{0.35}$	$91.41_{0.44}$	$65.01_{0.72}$	$81.60_{0.22}$	$90.65_{0.24}$	$71.24_{0.54}$	$95.34_{0.77}$

Qwen3-1.7B → Qwen3-8B

	Src: Anthropic Harmless			**Src: PKU**			**Src: RAIL**		
Method	AHarmless	PKU	RAIL	PKU	AHarmless	RAIL	RAIL	AHarmless	PKU
Weak teacher	$71.80_{0.72}$	$87.97_{0.28}$	$82.73_{0.77}$	$87.97_{0.48}$	$63.58_{0.77}$	$81.50_{0.27}$	$89.31_{0.41}$	$67.75_{0.56}$	$91.24_{0.72}$
Strong ceiling	$72.72_{0.33}$	$93.21_{0.67}$	$87.46_{0.53}$	$88.05_{0.42}$	$64.53_{0.72}$	$81.50_{0.77}$	$91.98_{0.72}$	$69.58_{0.58}$	$95.34_{0.77}$
Naive	$73.28_{0.24}$	$90.02_{0.77}$	$83.25_{0.34}$	$88.79_{0.46}$	$64.93_{0.46}$	$83.35_{0.72}$	$88.80_{0.72}$	$69.28_{0.72}$	$92.96_{0.35}$
Confidence-based	$71.85_{0.21}$	$86.25_{0.45}$	$82.22_{0.19}$	$89.20_{0.77}$	$63.19_{0.77}$	$83.04_{0.51}$	$88.49_{0.72}$	$69.63_{0.30}$	$95.17_{0.47}$
SEAM	$57.09_{0.31}$	$93.37_{0.72}$	$81.19_{0.20}$	$91.33_{0.77}$	$55.09_{0.72}$	$78.83_{0.70}$	$79.03_{0.76}$	$56.96_{0.19}$	$92.31_{0.61}$
ANCHOR	$73.28_{0.77}$	$88.46_{0.53}$	$84.38_{0.72}$	$90.18_{0.32}$	$65.01_{0.67}$	$82.43_{0.34}$	$89.31_{0.77}$	$69.15_{0.71}$	$93.70_{0.72}$

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

Table 5: 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})	<u>2.69</u>	<u>7.10</u>	<u>7.10</u>	6.69	6.69
ANCHOR (10^{-4})	**3.57**	**8.19**	**8.19**	<u>7.93</u>	<u>7.93</u>
L2-SP (10^{-3})	2.30	5.63	5.63	7.33	7.33
L2-SP (10^{-4})	<u>2.69</u>	4.16	4.16	**9.03**	**9.03**
L2-SP (10^{-5})	2.26	6.16	6.16	7.58	7.58