ALIGNING PREDICTION RESIDUALS: REPRESENTATION-AWARE DATA SELECTION FOR WEAK-TO-STRONG GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weak-to-strong generalization (W2S), where a weaker teacher supervises the training of a more capable student, offers a path to advancing frontier models beyond what costly human expert supervision can offer. Existing W2S work focuses on model-centric aspects such as adapting the training objective. Far less is known about which data to use in order to maximize the student's improvement over its teacher. We study this problem, beginning with a characterization of how bias and structured noise in weak labels limit the performance of the student. Building on this, we develop a data selection framework, **RADS**, measuring differences in predictive capability between the two models' representations. RADS prioritizes points with features the stronger model encodes but the weaker teacher misses, and training on them yields the largest performance gains. Across nine datasets at a 50% selection budget, RADS achieves a mean performance gap recovered (PGR) of 0.62, exceeding the strongest non-RADS baseline by 0.1. We show this advantage does not come from simply selecting samples the teacher already labels correctly, but from data with features the student can additionally learn from. For reward modeling, RADS exceeds random selection in mean PGR by 0.06 and 0.101 at 10% and 30% selection budgets, respectively. Finally, we present a probing study that estimates potential W2S gains prior to training [1].

1 INTRODUCTION

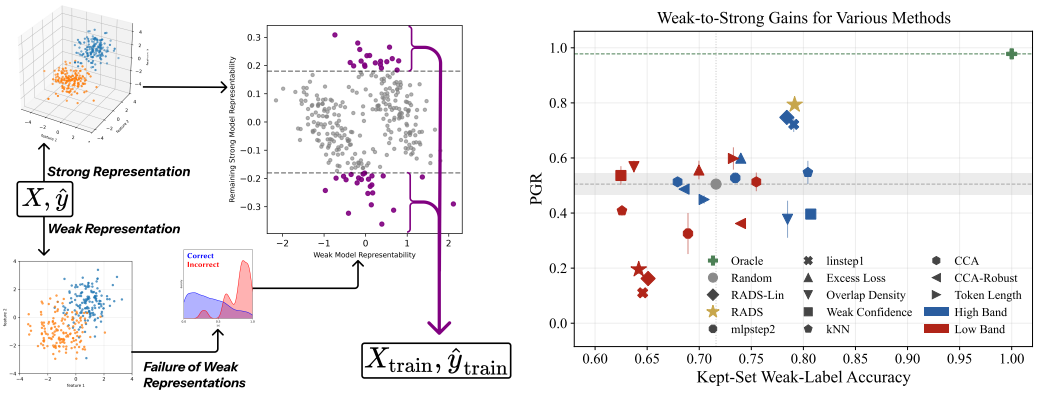
Weak-to-strong generalization (W2S) is a counterintuitive phenomenon where a stronger student model, trained on the imperfect supervision of a weaker teacher, nonetheless outperforms that teacher (Burns et al., 2024). As models approach superintelligence, their growing advantage over humans in speed and skill puts frontier systems beyond what human supervision can readily help train, making it increasingly necessary for existing models (e.g., previous generation) to train and audit the more capable models that succeed them. This need is further strengthened by the notion of recursive self-improvement (Fu et al., 2013; Moser et al., 2025; Duan et al., 2026). However, such model-to-model iteration can be inefficient, as noisy weak supervision may mislead the strong model; worse, the strong model may inherit its teacher's misalignment (Ji et al., 2025).

Training data is the key ingredient for transferring capabilities from the weak model to the strong one. Prior work has studied properties of data that enable maximizing weak-to-strong generalization (Shin et al., 2025; Xue et al., 2025). When the data lacks these properties (e.g., the presence of multiple types of feature patterns), the strong model performs no better than its weak teacher. Data selection therefore offers a practical way to improve W2S generalization, in terms of both efficiency and performance. Removing points that carry little signal for bridging the gap between weak and strong representations leaves the student with a cleaner training signal from which to generalize.

Existing approaches seeking to perform weak-to-strong data selection, however, fall short in two respects: they **(i)** may not perform well when going beyond linear models, and **(ii)** lack fine-grained control over data prioritization. First, several methods are developed in simplified, often linear, settings (Mulgund & Pabbaraju, 2025; Somerstep et al., 2025), which permit tractable theoretical

[1] Code is available at https://anonymous.4open.science/r/rads-w2s-generalization-78A1.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) How RADS scores are determined.

(b) ANLI-R1 under a 50% data budget.

Figure 1: **RADS selects data points that drive larger weak-to-strong gains. (a)** RADS scores each point by how much information the strong model's representation captures about it that the weak model's does not; higher-scoring points are expected to yield larger W2S gains. **(b)** Selecting data this way pays off in practice: for example, on ANLI-R1 with a 50% data budget, RADS (filled gold star) achieves the highest PGR of any method other than training on oracle labels.

analysis but offer no guarantee of transfer to realistic scenarios. This line of work also identifies dataset-level metrics that predict W2S gains. Optimizing such a metric through naive sampling risks violating the i.i.d. assumption it relies on, since the selected samples may not be representative of the full dataset (Xue et al., 2025; Dong et al., 2025). Second, methods designed for realistic settings select data at a coarse granularity: each sample is chosen via a binary keep-or-discard decision (Shin et al., 2025). Without continuous scores, samples cannot be prioritized precisely.

To overcome these, we introduce **RADS**, a two-stage data selection framework. First, we extract the teacher's representations and fit an MLP on them to predict the weak labels. The residual of this fit isolates the part of each label that the weak representation cannot explain. Second, we use kernel ridge regression on the student's representations and apply it to these residuals, discarding whatever the strong representation cannot explain either. The projection yields a per-sample score, measuring how much of the label the strong model captures but the weak model misses. Building on this, we obtain a fine-grained criterion for filtering the training set so the student learns more effectively.

We evaluate RADS and its variants on nine datasets against a broad range of data filtering strategies from the W2S literature. Using a 0.5B teacher and a 7B student, RADS improves PGR over random selection by 0.14, whereas the strongest prior method gains only 0.04 in PGR. For instance, Figure 1b shows that on ANLI-R1 (Nie et al., 2020), RADS yields the greatest W2S gain of 0.79, even though its filtered data has lower weak label accuracy. The improvement holds across data budgets: on ANLI-R1 and SST-2 (Socher et al., 2013), RADS frequently outperforms weak confidence and the next strongest method, especially at medium budgets between 40% and 60% selection. Extending to reward model training, a common component of recursive self-improvement, RADS is the only tested non-oracle method that outperforms random sampling on average at both tested budgets. Finally, we propose a probing metric that predicts how much RADS gains over training on the remaining, unselected points; this probing metric correlates with the observed gain at $r = 0.84$.

We summarize our contributions as follows:

- We introduce RADS, a representation-alignment data selection method for improving W2S training, without access to any ground-truth labels. RADS yields the largest W2S gains among all the baselines we compare against.
- We show these gains hold across data budgets and extend beyond classification to reward modeling, where RADS is the only non-oracle method that outperforms random sampling on average.
- We identify a training-free metric that predicts the PGR difference between training on the selected and discarded halves.

2 Related Work

Weak-to-strong (W2S) generalization is the phenomenon where a strong student, trained on a weak teacher's predictions, outperforms the teacher (Burns et al., 2024). This notion raises two questions: how can it be improved? and what causes W2S generalization? We group prior work into **(i)** empirical methods and **(ii)** theoretical analyses.

Empirical Methods. One way to strengthen W2S generalization is to improve the training data, e.g., by filtering it. Several methods assume access to ground-truth labels and use them to train auxiliary models that make filtering decisions (Agrawal et al., 2025; Zhou et al., 2025). Others bypass ground truth by smoothing each point's label using its neighbors in an embedding space (Lang et al., 2025). Smoothing helps to identify a mislabeled point whose neighbors mostly carry the correct label, but it cannot fix structured errors, where an entire group of nearby points shares the same incorrect prediction. A third line of work relies on generation: prompting a strong model to reason about which data to keep (Li et al., 2024), or calibrating the weak model's confidence using consistency across multiple chain-of-thought samples (Shi et al., 2025; Guo & Yang, 2024). However, these methods require models that can reason and generate reliable justifications. In contrast, our method requires no additional ground-truth labels for selection. Closest to our work, Shin et al. (2025) identifies "overlap density" in model representations, a property of the data that leads to better W2S generalization. While effective, this metric is coarse-grained: it offers little flexibility to grow or shrink the selected subset, which, as we will show, leads to suboptimal performance.

Theoretical Analysis. In contrast, these analyses offer stronger guarantees in simplified settings. Many results propose plausible causes of W2S generalization (Mulgund & Pabbaraju, 2025; Somerstep et al., 2025). For example, Dong et al. (2025) identify settings in which low intrinsic dimension enables W2S generalization. Xue et al. (2025) analyze ridge regression and derive a per-dataset metric that predicts the performance gap between a W2S model and a strong model trained on ground truth labels. Similarly, Liu et al. (2026) shows that spurious correlations between weak and strong features lead to W2S generalization. However, these analyses are limited to simplified settings such as linear models, and it remains unclear whether their conclusions hold for other cases.

3 Data Selection Framework for W2S Generalization

In this section, we detail our two-stage data selection framework. We begin with our setup and a framework overview (Sec. 3.1). We follow with a data-filtering method using a dataset-level quality measure in Xue et al. (2025), titled RADS-Lin (Sec. 3.2). From this initial ridge regression-based method, we explore non-linear alternatives that is expected to be more conducive to language data (more complex settings), establishing the non-linear variant, RADS (Sec. 3.3). Moreover, we present an elementary setting where RADS improves upon no-filtering.

3.1 Setup and Overview

In the W2S setting, we are given a weaker model θ_w, a higher-capacity "stronger" model θ_s to be trained via signal from θ_w, and a dataset $\mathcal{D}$. The core question is: *how can the performance of the stronger model be increased while using only labels generated by the weaker model?* We address this by scoring each training sample, ruling out those that contribute less to improving generalization.

For each data point X_i, we obtain representations from both models, $(X_w)_i$ and $(X_s)_i$, and weak labels $\hat{\mathbf{y}}$. Our method operates with two steps: first, remove the components of these labels explained by the weak representations; then, retain the components explainable by the strong representations. What remains is a score used to filter the training samples. We begin by deriving RADS, inspired by prior theoretical work by Xue et al. (2025). The resulting method is detailed in Algorithm 1. We then refine this initial linear method to include nonlinear alignment, which is more suitable for language data. A complete description of all notation is provided in Appendix A.

3.2 Representations as Selection

We first derive the structure of RADS-Lin, the data selection method filtering through linear representation alignment. This method is inspired by the theoretical analysis of ridge regression done in Xue et al. (2025). This work established a bound on weak-to-strong performance based on the

Algorithm 1 Representation-Aware Data Selection (RADS)

Require: weak model θ_w, strong model θ_s, dataset $\mathcal{D}$ of size n, ridge coefficient λ
Ensure: subset $\mathcal{D}_w \subseteq \mathcal{D}$ for training the strong model

$X_w, X_s \leftarrow$ centered representations of $\mathcal{D}$ induced by θ_w, θ_s
$\hat{\mathbf{y}}_w \leftarrow \theta_w(\mathcal{D})$ ▷ *weak supervision*
Step 1: residual beyond the weak model's capacity
fit an MLP ϕ such that $\phi(X_w) \approx \hat{\mathbf{y}}_w$
$\mathbf{a} \leftarrow \hat{\mathbf{y}}_w - \phi(X_w)$ ▷ *unexplained signal*
Step 2: capacity of the strong representations
$K_s \leftarrow X_s X_s^\top$ ▷ *linear kernel*
$P_s \leftarrow K_s\big(K_s + \lambda \frac{\mathrm{tr}(K_s)}{n} I\big)^{-1}$ ▷ *ridge smoother*
$\mathbf{v} \leftarrow P_s \mathbf{a}$ ▷ *residual recoverable by θ_s*
$m \leftarrow \mathrm{median}\big(|\mathbf{v}|\big)$
return $\mathcal{D}_w = \big\{\, x \in \mathcal{D} \;:\; |\mathbf{v}_x| \geq m \,\big\}$

representation spaces of each of the strong and weak models. A core theorem from this work is: **Theorem 3.8** (informal, Xue et al. (2025)). *Assuming X_w is $(\delta_w, \hat{\gamma}_w, \tilde{\gamma}_w)$-decomposable, X_s is $(\delta_s, \hat{\gamma}_s, \tilde{\gamma}_s)$-decomposable, and regularization satisfies $\beta_w > \delta_w$ and $\beta_s > \delta_s$, then w.h.p., we have*

$$\textbf{PredGap} = \frac{1}{n}\|P_s(I - P_w)\mathbf{y}\|^2 \pm o(1).$$

We focus on the RHS of this main expression; the precise definition of $(\delta, \hat{\gamma}, \tilde{\gamma})$-decomposability is found in Xue et al. (2025). In this theorem, P_s and P_w represent softened projections that remove all components not explained by the strong or weak model, respectively. The components of the vector $\mathbf{y}$ are the ground-truth labels. At a high level, this theorem states that the gap between the weak-to-strong model and a strong model trained on perfect labels is asymptotically controlled by $P_s(I - P_w)\mathbf{y}$. We use this and split the calculation into two steps:

- **Step 1.** Identify the components of the ground-truth labels $\mathbf{y}$ not explained by the weak representations. This gives us $\mathbf{a} := (I - P_w)\mathbf{y}$.
- **Step 2.** Calculate the ability of the strong model to capture what the weak model could not. From this, we have $\mathbf{v} := P_s\mathbf{a} = P_s(I - P_w)\mathbf{y}$.

The value of $\|\mathbf{v}\|^2 = \sum_i |\mathbf{v}_i|^2$ is what controls **PredGap** in ridge regression. This suggests using $|\mathbf{v}_i|$ as a per-sample score: each component $|\mathbf{v}_i|$ captures the inability of the weak model to represent the data point x_i, yet the strong model can represent it. This motivates selecting samples with the highest $|\mathbf{v}_i|$. The theorem provides a motivation for the score; we further assess the benefits of high-score selection empirically. However, this calculation requires the ground-truth labels $\mathbf{y}$. We substitute these ground-truth labels with the weak labels $\hat{\mathbf{y}}$, scoring data with $\mathbf{v} = P_s(I - P_w)\hat{\mathbf{y}}$. Although the weak labels are noisy, we find this substitution effective empirically.

3.3 Non-Linear Structure

In RADS-Lin, both steps use a projection from ridge regression to determine the quality of representations. However, these projectors assume linearly decodable representations, unlike the nonlinear structure of language models and their representations. We test replacing these steps with nonlinear counterparts, augmenting each alignment by instead training a small MLP to map the features (X_w for step 1, X_s for step 2) to the target of that step ($\hat{\mathbf{y}}$ for step 1, $\mathbf{a}$ for step 2). The outputs of these MLPs serve as the scores for each step. We found that the best performance comes from using an MLP for Step 1 while using the ridge soft projector P_s for Step 2 (see Section 4). This algorithm, named RADS (Representation-Aware Data Selection), is described in Algorithm 1.

Illustrative Example. We expect RADS to produce W2S gains when it selects points with large $|\mathbf{v}|$. We use a simple example to illustrate the intuition behind it. We first consider a setting where both the weak and strong features exist in $\mathbb{R}^2$. For each point in the dataset $X_i \in \mathcal{D}$, draw i.i.d. a label $y_i \in \mathrm{Unif}(\{-1, 1\})$ and a noise vector $\xi_i \sim \mathcal{N}(0, \sigma^2 I_2)$. We define a variance scalar $S > 1$, and two mean vectors $\mu_s^+, \mu_w^+ \in \mathbb{R}^2$. Both mean vectors represent the means of $X_s \,|\, y = 1$ and $X_w \,|\, y = 1$.

Note that these do not have to be collinear; in fact, RADS outperforms random sampling precisely when they are not. The scalar S scales the associated noise for the weak model, making the labels noisier. The features are set as

$$(X_w)_i = y\mu_w^+ + S\xi_i, \qquad (X_s)_i = y\mu_s^+ + \xi_i.$$

Each label creates a distinct distribution, symmetric across the origin, with the noise scaled up for the weak representations.

Two properties of this setting make RADS more effective than no filtering. *First*, since $S > 1$, the weak features are noisier than the strong ones, which leaves room for the strong model to improve on the weak labels. *Second*, because both feature sets share the same noise ξ_i, the weak model's errors are *structured*: it mislabels groups of nearby points in the strong feature space rather than a random subset of the data. Such structured errors are what distinguish W2S training from training with i.i.d. label noise, and they make training on the full, unfiltered data suboptimal. Without filtering, a learner trained on the strong features with weak labels is systematically misled by these errors.

For example, with $\mu_s^+ = [1, 1]^\top$, $\mu_w^+ = [1, 0]^\top$, $S = 4$, and $\sigma^2 = 1$, RADS achieves an average PGR of 0.68 across 1000 seeds, compared to 0.56 without filtering. Notably, the weak-label accuracy on the selected data is not significantly different from that of the unfiltered data; thus, RADS uses a structure beyond label accuracy. Both behaviors match what we observe on real datasets.

4 Experiments

We evaluate RADS across a range of settings and organize the results into five claims.

- **C1.** Under a fixed data budget, RADS is the best-performing method on average across nine datasets (Sec. 4.1).
- **C2.** This advantage holds across different data budgets, closing most of the gap to a student model trained on ground-truth labels (Sec. 4.2).
- **C3.** The gains do not come solely from selecting more accurately labeled data: RADS also outperforms training on randomly denoised data (Sec. 4.3).
- **C4.** On reward modeling, RADS is the only non-oracle method that improves over random selection on average at both tested budgets (Sec. 4.4).
- **C5.** The intermediate steps of RADS exhibit structure that enables us to develop a representation-based metric for estimating W2S gains (Sec. 4.5).

Baseline Methods. We compare RADS and RADS-Lin against *seven* families of baselines from the literature, each designed to select samples that improve W2S performance. We focus on ground-truth-free methods, excluding those that require a held-out set of ground-truth labels to assess the quality of weak predictions (Agrawal et al., 2025; Zhou et al., 2025). Except for random selection, every baseline assigns a score to each training point, and we keep the top-scoring samples up to the data budget. The baselines are:

(i) **Random selection**, our reference point for improvement, which uniformly subsamples the weakly labeled data to the size used by the other methods.
(ii) **Weak confidence**, which scores each point by the weak model's predicted probability for its weak label, pruning points where the weak model is likely wrong.
(iii) **Excess loss**, which scores each point by the untuned strong model's loss minus the weak model's loss, measuring how much the strong model can still learn from it.
(iv) **CCA**, which fits a linear alignment between the weak and strong representations via canonical correlation analysis, and scores each point by the distance between its projected strong representation and its weak representation.
(v) **kNN**, which predicts whether a point's weak label is correct from its nearest neighbors in the strong representation space, and scores the point by that prediction.
(vi) **Token length**, which scores each point by its length in tokens, a proxy for sample difficulty.
(vii) **Overlap density** (Shin et al., 2025), which selects points that have high weak confidence and lie close to points with low weak confidence.

Evaluation Metric. We report the Performance Gap Recovered (PGR) metric of Burns et al. (2024),

$$\textbf{PGR} = \frac{\text{weak-to-strong} - \text{fine-tuned weak}}{\text{ground-truth strong} - \text{fine-tuned weak}},$$

Method	AG News	IMDb	SciQ	SST-2	ANLI-R1	FEVER-NLI	WANLI	HellaSwag	ETHICS	Avg
RADS	0.36±.08	**0.79**±.02	0.97±.01	0.79±.02	**0.79**±.01	**0.32**±.00	**0.62**±.02	0.58±.01	**0.35**±.02	**0.62**
RADS-Lin	0.56±.05	0.75±.01	**0.98**±.00	0.77±.07	0.75±.00	0.19±.02	0.46±.09	0.49±.01	0.33±.03	0.59
Excess Loss	**0.65**±.02	0.75±.01	0.97±.01	0.73±.02	0.60±.02	-0.04±.02	0.34±.05	0.48±.03	0.18±.02	0.52
CCA-Robust	0.46±.07	0.79±.02	0.96±.00	0.77±.04	0.49±.01	-0.00±.06	0.22±.00	0.48±.02	0.20±.01	0.48
Random Selection	0.41±.08	0.77±.03	0.97±.01	0.74±.05	0.51±.04	0.04±.05	0.24±.04	0.44±.07	0.24±.02	0.48
Overlap Density	0.43±.09	0.71±.02	0.97±.01	0.73±.02	0.38±.07	-0.02±.01	0.22±.05	0.59±.03	0.22±.01	0.47
Weak Confidence	0.26±.16	0.76±.04	0.95±.00	0.61±.09	0.40±.00	0.04±.01	0.29±.01	**0.63**±.05	0.23±.02	0.46
Token Length	0.27±.14	0.68±.02	0.96±.01	0.68±.01	0.45±.01	0.12±.03	0.42±.03	0.50±.03	0.06±.00	0.46
kNN	0.51±.07	0.62±.03	0.96±.00	0.58±.05	0.55±.04	-0.07±.04	0.20±.02	0.55±.05	0.21±.02	0.46
CCA	0.19±.04	0.72±.05	0.93±.02	**0.82**±.02	0.51±.02	-0.03±.02	0.19±.01	0.54±.09	0.20±.01	0.45
Base (Untuned Strong)	-2.21	0.54	0.92	0.42	-0.09	-0.04	-0.71	0.35	-0.02	-0.09

Table 1: Comparison of RADS to other competing methods. RADS and other variants of RADS perform on average better than the other baselines tested. The best entry of each column is in bold. The untuned strong row represents the PGR if the untuned strong performance is used. The sign of this value indicates if the untuned model outperforms the tuned weak model.

which measures how much of the gap between the fine-tuned weak model and a strong model fine-tuned on error-free labels is recovered by training the strong model on weak labels. A PGR of 1 means the gap is fully closed with strong model performance near training on ground-truth labels, and a PGR of 0 means weak supervision produces nothing over the weak model itself.

4.1 Comparison to Other W2S Filtering Methods

We compare RADS to a range of alternative selection methods across nine datasets. RADS yields the largest W2S gains of all methods considered.

Setup. We first use Qwen2.5-0.5B and Qwen2.5-7B as our weak and strong models, respectively (Qwen Team, 2025). Using base models from the same family may help ensure that their features are aligned. The distinction between the weak and the strong model comes solely from model size (i.e., knowledge capacity) rather than from performance; that said, we do observe that performance within our model pairs is often correlated.

We first collect nine natural language classification tasks (see dataset description in Appendix B.2). On each dataset, we first build the weak model to a classifier on 3,000 questions by fine-tuning the top layer using linear probing. This fine-tuned model gives the weak labels $\hat{\mathbf{y}}$ on another 3,000 data points (our training set), on which the strong model is then fine-tuned; this is the step where each method filters the data. We use the remainder as held-out testing set, depending on the dataset, to measure the PGR (Appendix B.1). In this experiment, we set the selection ratio to be 50% and run 3 trials with different seeds. For every method and dataset we report the PGR of the fine-tuned strong model, along with the per-method average. We additionally report the PGR if we used the base strong model. This is used to identify settings where the base strong model is already better than the fine-tuned weak models. These are indicated by a positive untuned strong PGR.

Results. Table 1 shows our main results, where methods are sorted by their average performance across the nine datasets. RADS and RADS-Lin are the two strongest methods, and RADS outperforms RADS-Lin on almost every dataset. RADS and RADS-Lin outperform the random sampling baseline by $+0.14$ and $+0.11$ in mean PGR, respectively. The next strongest method, which filters samples by measuring their learnability via excess loss, shows a separation of 0.1 in PGR.

Note that on many of these datasets, the base strong model already outperforms the fine-tuned weak model. In some of these cases, we suspect that the strong model has been pre-trained on the dataset itself, as appears to be the case for SciQ. Focusing on the datasets where the fine-tuned weak model outperforms the strong model (i.e., the untuned strong PGR is negative), RADS is the best-performing method on four of the five. Moreover, weak-confidence sampling performs surprisingly poorly across datasets. This highlights the risk of trusting the weak model's confidence: the smaller model does not judge its own ability well, which yields noisier labels. This applies more broadly to any setting that relies on an auxiliary model to guide data selection. Finally, Table 5 reports the PGR of models trained on the remaining 50% of the data, confirming that the samples RADS and RADS-Lin discard are indeed the least useful ones.

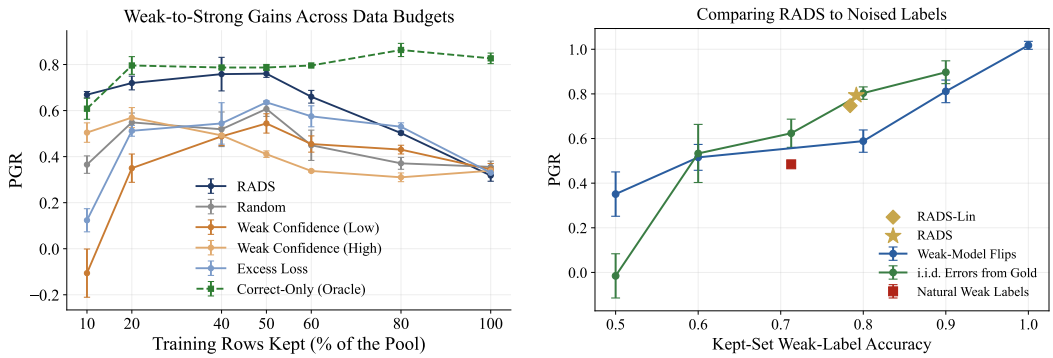


Figure 2: Even when changing the selection budget on ANLI-R1, RADS outperforms compared methods on PGR. Critically, this holds for low sample sizes. Additionally, RADS frequetly outperforms methods that denoise the weak labels in an i.i.d. manner.

4.2 Testing Other Selection Ratios

Sec. 4.1 fixes the selection ratio at 50%. This choice trades off the two failure modes of selection: a smaller ratio leaves too little data to train on, while a larger one admits points that hurt W2S performance. It also splits the data into two equal halves, which makes the comparison against training on the discarded points a fair one. However, we also aim for these methods to work well on other sampling scales when training is resource-constrained yet the highest possible PGR is desired.

Setup. We rank points by our RADS scoring metric $|\mathbf{v}|$ and keep the highest-scoring ones at each selection ratio. We use ANLI-R1 as our example dataset to match the results shown in Figure 1b. We compare against random selection, weak-confidence scoring, and the excess-loss method, and additionally include a correct-only oracle—which trains only on points whose weak labels match the ground-truth labels—as an upper bound. Since ground-truth labels are unavailable in practice, this oracle serves only as a reference.

Results. Figure 2 shows that RADS closes the gap to the upper bound the most among all selection strategies. In the low-budget regime, RADS is competitive with the oracle: ***selecting a small number of points by*** $|\mathbf{v}|$ ***recovers most of the attainable gain without access to ground-truth labels.*** As the budget grows, increasingly many noisy labels enter the training set and performance degrades for every method, yet RADS remains the strongest. Results for the other datasets are presented in the Appendix C.7.

4.3 Comparison to Noised Data

Figure 1b shows that weak-label accuracy may not be the best predictor of W2S gains. On ANLI-R1, RADS attains the highest PGR, even though methods such as weak-confidence scoring retains weak labels with higher accuracy. To understand why, we compare RADS and RADS-Lin against *two noising schemes*: one that preserves the bias in the weak labels and one that does not.

Setup. The two noising schemes come from different source labels. The first is i.i.d. corruption of the ground-truth labels, similar to other non-W2S settings. The other is a corruption centered at the weak labels: for error ratios above the weak labels, additional correct labels are flipped; for error ratios less than the weak labels, incorrectly labeled points have their labels changed to be correct. This second scheme is designed to maintain the biases inherent to the weak labels. We expect the second scheme to lead to worse performance due to structured bias in these labels.

Results. The results are shown in Figure 2. Both RADS and RADS-Lin attain a performance similar to the i.i.d. noised labels at the same weak-label accuracy. Notably, these methods surpass the setting where the labels are uncorrupted from the weak labels. Thus, the data selection done by these algorithms goes beyond what i.i.d. denoising of the weak labels is capable of, instead selecting data more similar to the ground truth with noise.

Table 2: PGR (AUC) for reward modeling at labeling budgets of 10% and 30%. RADS is the only method that outperforms random selection on average at both budgets.

Method	Dataset				
	JudgeLM	**Preference 700K**	**Prometheus**	**UltraFeedback**	**Avg**
Budget = 0.1					
Correct-Only	1.113	1.419	1.715	1.273	1.380
RADS	**0.188**	0.328	0.200	**0.789**	**0.376**
Random Selection	0.079	**0.628**	**0.226**	0.333	0.316
Token Length	−0.045	0.318	−0.059	0.789	0.251
Weak Confidence	−0.804	0.500	−0.857	0.441	−0.180
Budget = 0.3					
Correct-Only	0.849	0.978	1.022	0.947	0.949
RADS	**0.409**	0.299	0.506	**0.523**	**0.434**
Token Length	0.218	0.438	**0.524**	0.500	0.420
Random Selection	0.199	0.408	0.331	0.393	0.333
Weak Confidence	0.030	**0.470**	−0.192	0.422	0.182

4.4 Reward Modeling

Our experiments so far (Sec. 4.1) focus on NLP classification. We further explore whether RADS's advantage carries over to reward modeling, a key component of modern self-improvement pipelines.

Setup. We employ four widely used preference datasets: JudgeLM (Zhu et al., 2025), Preference 700K (Dong et al., 2024), Prometheus (Kim et al., 2024), and UltraFeedback (Cui et al., 2023). From each dataset, we sample 10,000 examples and split them 70/10/20 into training, validation, and test sets. We first fine-tune Qwen2.5-3B-Instruct on the validation set to obtain the weak reward model as teacher, and use it to label the training set with its predicted preferences. Given a data budget of 10% or 30%, each method selects the top-ranked weakly labeled points, on which we fine-tune Qwen2.5-7B-Instruct as the strong student. We measure the student's test AUC and report the resulting PGR. We compare RADS against three baselines: random selection, token length, and weak-model confidence. We also include an oracle reference, *correct-only*, which uses ground-truth labels to select only the examples the weak teacher labeled correctly.

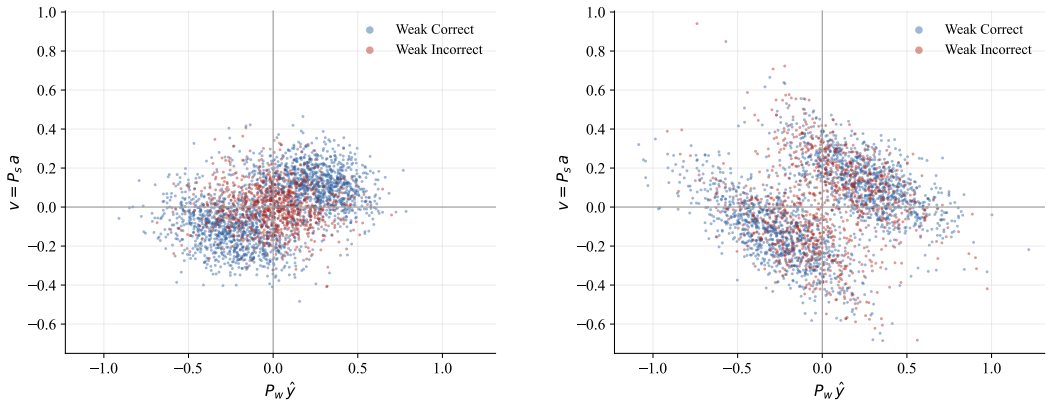
Results. Table 2 shows how data selection affects W2S generalization in the reward modeling setting. RADS achieves the highest average PGR at two budgets (0.376 at 10% and 0.434 at 30%), and it is the only method that outperforms random selection on average at both. It thus narrows the gap to the correct-only reference, which requires oracle access to ground-truth labels. Consistent with Sec. 4.1, selection by weak-model confidence is among the least effective strategies.

4.5 Predicting PGR without Training

So far, RADS has shown remarkable ability to perform well on a range of datasets. However, some datasets exhibit the opposite behavior: taking points with high $|\mathbf{v}|$ results in significantly worse performance, while using points with low $|\mathbf{v}|$ results in the strongest performance. We now present a method used before fine-tuning the strong model to predict which band yields higher performance.

Setup. We use the strong and weak representations to predict the difference in performance between selecting high $|\mathbf{v}|$ and low $|\mathbf{v}|$. The statistic of interest is the centered kernel alignment (CKA), previously used to determine the similarity between neural representations (Kornblith et al., 2019). We expect a relationship between the similarity between the representations and the performance of RADS, as this method relies upon structure in the representations of the strong and weak models.

We also show the relationship between $\hat{\mathbf{y}} - \mathbf{a}$ and $\mathbf{v}$ for two datasets: ANLI-R1 and SciQ. These are selected to show the two distinct behaviors observed across the datasets. Within RADS, $\mathbf{a}$ captures the aspects of $\hat{\mathbf{y}}$ that the weak model cannot capture, while $\mathbf{v}$ retains the component explainable by the strong representations. We show the same correlations for other datasets in Figure 5.



(a) Correlation results for ANLI-R1 data.

(b) Correlation results for SciQ data.

Figure 3: Example datasets showing weak representation capacity and additional strong representation capacity. ANLI-R1 shows a correlation, with correct data lying disproportionately away from the origin. SciQ shows almost no correlation and a distinct structure from ANLI-R1. Correlation is predictive of PGR.

Results. The CKA for 23 datasets and associated RADS PGR differences are shown in Figure 4. The results show a Spearman correlation of 0.84 between CKA and high-band minus low-band PGR. In contrast, using weak label accuracy to predict PGR shows a correlation of 0.08. Therefore, shared structure between the two models is vital to performance, *not* weak label accuracy directly.

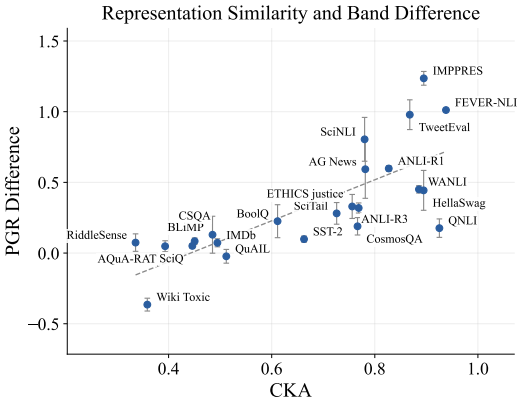


Figure 4: Correlation between CKA between strong and weak representations, and the PGR difference between RADS with high $|\mathbf{v}|$ and low $|\mathbf{v}|$.

For ANLI-R1 and SciQ, the relationship between $\hat{\mathbf{y}} - \mathbf{a}$ and $\mathbf{v}$ is shown in Figure 3. These two datasets show remarkably different structures. The distribution for ANLI-R1 is roughly unimodal, with accurate points far from the origin and accurate points near the origin. In contrast, the SciQ data showcases a strongly bimodal distribution. This bimodal distribution stems from the binary weak labels it was trained on, suggesting that the weak model is confident in its representations, and thus the strong model mirrors the weak model. Supporting this, the PGR above random on ANLI-R1 is significantly higher than on SciQ.

5 Conclusion

Recursive self-improvement is perhaps the key future paradigms for frontier model training. Performing it requires weak-to-strong generalization—where weaker teachers train stronger students to outperform them. We investigated W2S generalization from a data-centric perspective. Inspired by representation alignment known to determine W2S gains in ridge regression, we develop RADS; this method samples points where the strong model representations capture features that the weak model representations do not. We show that RADS outperforms its competitors while using no held-out ground-truth labels. This holds across data budgets and across both classification and reward modeling settings. Using the relationship between the student's and teacher's representations, we identified a metric that predicts the PGR difference between training on the selected and discarded portions of data while requiring no language model training.

AI USE STATEMENT

In this work, we used generative AI to help with the literature search, identifying relevant papers on W2S generalization and RSI. During writing, we used generative AI to assist with grammatical correctness. We additionally used generative AI tools to help write the code, clean and format datasets, interpret intermediate results, and create figures and tables. We did not use generative AI tools for any other required-disclosure task, including drafting sections of the paper or creating the illustrative setting. We have reviewed all AI-assisted work: AI-generated code was read, executed, and verified for correctness by 3 authors, and all suggested grammatical edits were checked to ensure no claims or technical content were altered. All final interpretations were made without the assistance of generative AI. Generative AI did not write any components of the related work or citations in the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Aakriti Agrawal, Mucong Ding, Zora Che, Chenghao Deng, Anirudh Satheesh, Bang An, Bayan Bruss, John Langford, and Furong Huang. Ensemw2s: Enhancing weak-to-strong generalization with large language model ensembles, 2025. URL https://arxiv.org/abs/2410.04571.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*, 2020.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ghNRg2mEgN.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. Toxic comment classification challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge, 2017. Kaggle.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, 2019.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

Yijun Dong, Yicheng Li, Yunai Li, Jason D. Lee, and Qi Lei. Discrepancies are virtue: Weak-to-strong generalization through lens of intrinsic dimension. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14079–14113. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/dong25g.html.

Yi Duan, Ying Liu, Zirui Tang, Haodong Chen, Jun Zhou, Yumou Liu, Bangrui Xu, Yukai Wu, Sidi Chen, Yuhan Zhou, et al. The last ai built by humans: Toward genuine recursive self-improvement. *arXiv preprint arXiv:2609.11873*, 2026.

Yifan Fu, Xingquan Zhu, and Bin Li. A survey on instance selection for active learning. *Knowledge and Information Systems*, 35(2):249–283, 2013.

Yue Guo and Yi Yang. Improving weak-to-strong generalization with reliability-aware alignment, 2024. URL https://arxiv.org/abs/2406.19032.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2391–2401, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1243. URL https://www.aclweb.org/anthology/D19-1243.

Paloma Jeretič, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8690–8705, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.768. URL https://www.aclweb.org/anthology/2020.acl-main.768.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Hua Xu, Aidan O'Gara, Kwan Ng, Brian Tse, Jie Fu, Stephen Mcaleer, Yanfeng Wang, Mingchuan Yang, Yunhuai Liu, Yizhou Wang, Song-Chun Zhu, Yike Guo, Yaodong Yang, and Wen Gao. Ai alignment: A contemporary survey. *ACM Comput. Surv.*, 58(5), November 2025. ISSN 0360-0300. doi: 10.1145/3770749. URL https://doi.org/10.1145/3770749.

Tushar Khot, Ashish Sabharwal, and Peter Clark. SciTail: A textual entailment dataset from science question answering. In *AAAI*, 2018.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8euJaTveKw.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Hao Lang, Fei Huang, and Yongbin Li. Selective weak-to-strong generalization, 2025. URL https://arxiv.org/abs/2511.14166.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14255–14273, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.769. URL https://aclanthology.org/2024.acl-long.769/.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. 2021.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the*

55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 158–167, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015/.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. Wanli: Worker and ai collaboration for natural language inference dataset creation, January 2022. URL https://arxiv.org/pdf/2201.05955.

Chenruo Liu, Yijun Dong, and Qi Lei. Does weak-to-strong generalization happen under spurious correlations? In *International Conference on Learning Representations*, volume 2026, pp. 32657–32687, 2026.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Brian B. Moser, Arundhati S. Shanbhag, Stanislav Frolov, Federico Raue, Joachim Folz, and Andreas Dengel. A coreset selection of coreset selection literature: Introduction and recent advances. *ArXiv*, abs/2505.17799, 2025. URL https://api.semanticscholar.org/CorpusID:278886355.

Abhijeet Mulgund and Chirag Pabbaraju. Relating misfit to gain in weak-to-strong generalization beyond the squared loss. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=wITGhlMPXz.

Yixin Nie, Haonan Chen, and Mohit Bansal. Combining fact extraction and verification with neural semantic matching networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01):6859–6866, 2019. doi: 10.1609/aaai.v33i01.33016859.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Qwen Team. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 8722–8731. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6398.

Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 502–518, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2088. URL https://aclanthology.org/S17-2088/.

Mobashir Sadat and Cornelia Caragea. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7399–7409, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.511. URL https://aclanthology.org/2022.acl-long.511/.

Junhao Shi, Qinyuan Cheng, Zhaoye Fei, Yining Zheng, Qipeng Guo, and Xipeng Qiu. How to mitigate overfitting in weak-to-strong generalization? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16100–16118, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.784. URL https://aclanthology.org/2025.acl-long.784/.

Changho Shin, John Cooper, and Frederic Sala. Weak-to-strong generalization through the data-centric lens. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uogG8BfLs2.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1170.

Seamus Somerstep, Felipe Maia Polo, Moulinath Banerjee, Yaacov Ritov, Mikhail Yurochkin, and Yuekai Sun. A transfer learning framework for weak to strong generalization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PeLLMw3wLX.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL https://doi.org/10.1162/tacl_a_00321.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

Yihao Xue, Jiping Li, and Baharan Mirzasoleiman. Representations shape weak-to-strong generalization: Theoretical insights and empirical predictions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ypEW077kle.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Yucheng Zhou, Jianbing Shen, and Yu Cheng. Weak to strong generalization for large language models with multi-capabilities. In *International Conference on Learning Representations*, volume 2025, pp. 11583–11612, 2025.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xsELpEPn4A.

Symbol	Meaning
X_w	The weak representations.
X_s	The strong representations.
K_w	The weak kernel, $K_w = X_w X_w^\top$
K_s	The strong kernel, $K_s = X_s X_s^\top$
P_w	The ridge-regression soft projector from X_w.
P_s	The ridge-regression soft projector from X_s.
ψ_w	The weak model.
ψ_s	The strong model.
ϕ	An MLP trained on the representations.
$\mathbf{y}$	The ground truth labels.
$\hat{\mathbf{y}}$	The weak labels.
$\mathbf{a}$	The weak labels with the components explained by the weak model removed.
$\mathbf{v}$	$\mathbf{a}$ with the components not explained by the strong model removed.

Table 3: Notation

A Notation

We provide a complete description of all notation used in Table 3.

B Experimental Details

The primary setting of W2S in this work is classification, as has been done previously (Burns et al., 2024). This section details the experimental choices to test W2S generalization.

B.1 Classification Tasks

Data Preprocessing. We use the classification datasets listed in Table 4, drawing the questions from the training split of each dataset. Each question contributes a single question-answer pair: with probability one half, the candidate is the correct answer (label 1), and otherwise it is one of the incorrect answers chosen at random (label 0). For most datasets, the pair takes the form "Q: ... A: <answer>" with the candidate in place of <answer>, and Appendix B.2 gives the format of each dataset in Table 1. We then subsample the pairs so that the two labels are equally frequent and split them into two equal halves. For the nine datasets in Table 1, each half has 3,000 questions.

Models. We use the Qwen2.5 family of models, with the 0.5B and 7B parameter models used as our weak and strong models, unless otherwise specified. The strong model training is done through PEFT of the larger Qwen model. This further stretches the capabilities of the strong and the weak model, creating a more optimal testbed for probing W2S generalization. Specifically, we fine-tune the strong model with LoRA (Hu et al., 2022) to answer "yes" or "no" to the question "Is the candidate answer correct?", which we append to each question-answer pair. RADS uses the last-layer representations of both models at the last token position, the same position as the linear probe of the weak model.

Training Hyperparameters. We fine-tune for 2 epochs with AdamW, a learning rate of 2×10^{-5}, 20 warmup steps followed by linear decay, an effective batch size of 4, gradient clipping at 1.0, and no weight decay. The LoRA adapters have rank 8, $\alpha = 16$, and dropout 0.05, and are applied to all attention and MLP projections. When fine-tuning and evaluating the strong model, we truncate inputs from the left to 512 tokens for IMDb and to 384 tokens for the other datasets in Table 1.

Weak Labels. To extract weak labels, we provide the weak model with the question-answer pairs described above. The weak model is then used to determine if a given question-answer pair is correct or incorrect. The labels are calculated through a linear probe using the last-layer representations of the weak model at the last token position. We train this probe, a logistic regression, on the first half with ground-truth labels and use it to label the second half, which forms the pool that every selection method draws from. A pair is labeled correct when the probe assigns it a probability of at least 0.5.

Dataset Name	Source
AG News	Zhang et al. (2015)
IMDB	(Maas et al., 2011)
SciQ	Welbl et al. (2017)
SST-2	Socher et al. (2013)
ANLI	Nie et al. (2020)
FEVER-NLI	Thorne et al. (2018); Nie et al. (2019)
WANLI	Liu et al. (2022)
HellaSwag	Zellers et al. (2019)
ETHICS	Hendrycks et al. (2021)
RiddleSense	Lin et al. (2021)
AQuA-RAT	Ling et al. (2017)
Wiki Toxic	cjadams et al. (2017)
BLiMP	Warstadt et al. (2020)
CSQA	Talmor et al. (2019)
QuAIL	Rogers et al. (2020)
BoolQ	Clark et al. (2019)
SciTail	Khot et al. (2018)
SciNLI	Sadat & Caragea (2022)
CosmosQA	Huang et al. (2019)
QNLI	Wang et al. (2018); Rajpurkar et al. (2016)
TweetEval	Barbieri et al. (2020); Rosenthal et al. (2017)
IMPPRES	Jeretič et al. (2020)
JudgeLM	Zhu et al. (2025)
Preference 700K	Dong et al. (2024)
Prometheus	Kim et al. (2024)
UltraFeedback	Cui et al. (2023)

Table 4: Datasets and their sources

Each selection method keeps the half of the pool with the highest scores (the high band) or the lowest scores (the low band); Random Selection keeps a random half. The kept pairs are sampled per class separately so that the two weak labels are equally frequent. The strong model is then fine-tuned on the kept pairs, with their weak labels as targets.

Evaluation. The strong model labels are generated through prompting, selecting the top choice among the possible classes: we score every candidate answer of a test question by the probability that the strong model answers "yes", and the question counts as correct when the top candidate is the correct answer. For PGR (Section 4), the fine-tuned weak performance is the accuracy of the linear probe on the same test questions, and the ground-truth strong performance is that of a strong model trained on the whole pool with ground-truth labels. We report the mean PGR over three training seeds with its standard deviation. For Random Selection, we average over five random halves of the pool, each trained with three seeds, and report the standard deviation across the halves.

B.2 Dataset Descriptions

We provide some details for each of the nine datasets in Table 1.

AG News (Zhang et al., 2015). News articles labeled with one of four topics: world, sports, business, and science and technology. Each query is posed as "What is the topic of this article?" with the four topic names as the candidate answers.

IMDb (Maas et al., 2011). Movie reviews from IMDb labeled positive or negative. Each review is posed as "What is the sentiment of this review?" with the two labels as candidates.

SciQ (Welbl et al., 2017). Crowdsourced science exam questions with one correct answer and three distractors. The question is presented without the supporting passage, and the four answers are the candidates.

Method	AG News	IMDb	SciQ	SST-2	ANLI-R1	FEVER-NLI	WANLI	HellaSwag	ETHICS	Avg.
kNN	0.48±.02	0.78±.04	0.96±.01	**0.79**±.01	0.41±.02	0.08±.06	0.22±.06	0.40±.07	0.26±.02	**0.49**
Random Selection	0.41±.08	0.77±.03	0.97±.01	0.74±.05	0.51±.04	0.04±.05	0.24±.04	0.44±.07	0.24±.02	0.48
Token Length	**0.53**±.08	0.60±.04	0.96±.00	0.64±.05	**0.60**±.04	-0.04±.04	0.24±.02	0.41±.01	0.26±.04	0.47
CCA-Robust	0.35±.10	0.79±.05	0.97±.00	0.70±.03	0.36±.01	0.00±.05	0.24±.03	0.30±.05	0.23±.01	0.44
CCA	0.36±.08	0.63±.05	**0.97**±.00	0.74±.03	0.51±.03	-0.10±.05	0.22±.08	0.31±.01	0.22±.02	0.43
Weak Confidence	0.40±.03	**0.79**±.07	0.97±.00	0.71±.04	0.54±.03	-0.39±.02	**0.25**±.03	0.23±.05	0.26±.01	0.42
Excess Loss	-0.09±.09	0.50±.04	0.97±.01	0.64±.02	0.56±.03	**0.19**±.05	-0.09±.03	**0.59**±.01	**0.28**±.01	0.39
Overlap Density	0.21±.11	0.69±.03	0.95±.00	0.75±.01	0.57±.02	-0.04±.04	-0.52±.29	0.28±.04	0.24±.04	0.35
RADS-Lin	-0.21±.08	0.68±.07	0.95±.01	0.44±.03	0.16±.02	-0.63±.02	0.15±.02	0.52±.02	0.02±.03	0.23
RADS	-0.23±.19	0.72±.02	0.92±.01	0.69±.01	0.20±.02	-0.69±.00	0.17±.02	0.14±.14	0.03±.03	0.22
Base (Untuned Strong)	-2.21	0.54	0.92	0.42	-0.09	-0.04	-0.71	0.35	-0.02	-0.09

Table 5: Results for training on low-bands rather than high-bands. A method filtering the strongest points for W2S should achieve the lowest performance here.

SST-2 (Socher et al., 2013). Single sentences from movie reviews labeled positive or negative, in the binary version distributed with GLUE (Wang et al., 2018). Each sentence is queried with "What is the sentiment of this sentence?". The test labels are not public, so the test questions come from the validation split.

ANLI (Nie et al., 2020). Each premise is paired with a hypothesis that annotators wrote to fool a model in the loop. We use the first round. The candidates are the three relations: entailment, neutral, and contradiction, asked as "What is the relationship from the premise to the hypothesis?".

FEVER-NLI (Thorne et al., 2018; Nie et al., 2019). Claims from FEVER paired with textual evidence from Wikipedia, recast as natural language inference by Nie et al. (2019). The candidates are supported, refuted, and not enough info, asked as "Based on the evidence, the claim is?". The test split is unlabeled, so the test questions come from the development split.

WANLI (Liu et al., 2022). Premise-hypothesis pairs generated by GPT-3 from patterns found in MultiNLI, then revised and labeled by crowdworkers. The candidates and the question are the same as for ANLI.

HellaSwag (Zellers et al., 2019). Each context from ActivityNet or WikiHow is paired with four candidate endings, three of which are machine-generated and adversarially filtered. The context and a candidate ending are joined with no question. The test labels are not public, so the test questions come from the validation split.

ETHICS (Hendrycks et al., 2021). We use the Justice subset: first-person statements about what someone deserves or about treating people impartially, each labeled reasonable or unreasonable. Each prompt is posed as "Is this claim reasonable or unreasonable?".

C Supplementary Experiments

We detail additional experiments to examine the relationships among the different components of RADS. These include training on other sampling strategies, using a different pair of strong and weak models, more experiments on the structural results uncovered in the data, and a few miscellaneous experiments justifying design decisions in RADS.

C.1 Results for low bands

In the main body, all of the tested methods used their high bands, i.e., sampling points with the largest $|\mathbf{v}|$ scores. The choice of high $|\mathbf{v}|$ or low $|\mathbf{v}|$ can be justified in different situations. A low $|\mathbf{v}|$ decreases **PredGap** for the sampled data points but ignores the points ignored by sampling. This is not guaranteed. In many situations, and the ones we observe in this work, sampling points with high $|\mathbf{v}|$ selects points where the strong model can extrapolate structure in the weak features. This, however, requires that the weak label errors do not correlate in adversarial ways with the strong representations. In light of this lack of universal specification of the optimal choice, we perform experiments on both the low $|\mathbf{v}|$ and high $|\mathbf{v}|$ to determine which performs better.

Method	AG News	IMDb	SciQ	SST-2	ANLI-R1	FEVER-NLI	WANLI	HellaSwag	ETHICS	Avg.
RADS	0.36±.08	**0.79**±.02	0.97±.01	0.79±.02	**0.79**±.01	0.32±.00	**0.62**±.02	**0.58**±.01	**0.35**±.02	**0.62**
linstep1	**0.57**±.06	0.76±.00	**0.98**±.00	0.74±.03	0.72±.03	0.25±.03	0.52±.05	0.49±.07	0.31±.00	0.59
RADS-Lin	0.56±.05	0.75±.01	**0.98**±.00	0.77±.07	0.75±.00	0.19±.02	0.46±.09	0.49±.01	0.33±.03	0.59
mlpstep2	0.44±.06	0.75±.03	0.96±.00	**0.85**±.04	0.53±.04	**0.33**±.03	0.43±.04	0.11±.06	0.15±.04	0.51

Table 6: The performance of RADS and all variants. The highest-performing variant is the primary algorithm, RADS.

The main body contains the results for the high bands in Table 1. For all datasets and methods tested, we also test using points in the low bands. The results are found in Table 5. First, we observe that many of the competitive methods behave how we expect: selecting the low bands results in performances below a random subsampling of the data. For example, using high excess loss results in an average performance of 0.52, while using low excess loss results in an average performance of 0.39.

For RADS and RADS-Lin, the difference is even more drastic. For each, the average PGR on high bands is 0.62 and 0.59, while on low bands, the average PGR is 0.22 and 0.23. Across all methods, this is the greatest separation between the two splits of the data. Thus, RADS and its variants provide a stronger data scoring metric than their competitors.

C.2 RADS VARIANTS

We established RADS-Lin and its refinement RADS. Both of these methods follow the same structure: start with removing the explainability of the weak labels using the weak representations, then keep only what can be explained by the strong model. The difference in RADS that is not in RADS-Lin is that the first step uses a small cross-fitted MLP rather than a linear predictor to determine what can and cannot be represented with the features. The second step is left unchanged. We test two other variants as well: mlpstep2 and linstep1.

mlpstep2. In mlpstep2, the second steps of RADS-Lin is replaced by an MLP. Unlike RADS, the strong model representations can be parsed with a nonlinear structure rather than a purely linear.

linstep1. This method is identical to RADS, except the first step uses a linear predictor instead of the MLP, still cross-fit.

The results for all 4 methods, including RADS and RADS-Lin, are found in Table 6. From this table, we observe that RADS is indeed the highest-performing of all other design choices. This includes mlpstep2, indicating that an MLP decoding the strong representations leads to overfitting rather than increased W2S gains. As for linstep1, the decreased average performance shows that the MLP in the first step helps; linstep1 and RADS-Lin, which differ only in cross-fitting, perform about the same.

C.3 ALTERNATIVE REPRESENTATION-BASED METHODS: CCA AND KNN

We compare our results to two other linear methods using representations to identify mislabeled points.

CCA. Canonical correlation analysis establishes a linear relationship between the strong representations and the weak representations. Samples are chosen depending on how far apart the strong projection and the weak projection are in this optimal linear alignment.

This method relies on representations between stronger and weaker models being roughly linear. One modification of this method includes excluding outliers to make the alignment more robust.

kNN. The kNN classifier simply takes the nearest neighbors in the strong representation space and uses these to predict whether the weak label of the current point is correct. The prediction becomes the score to sample based on.

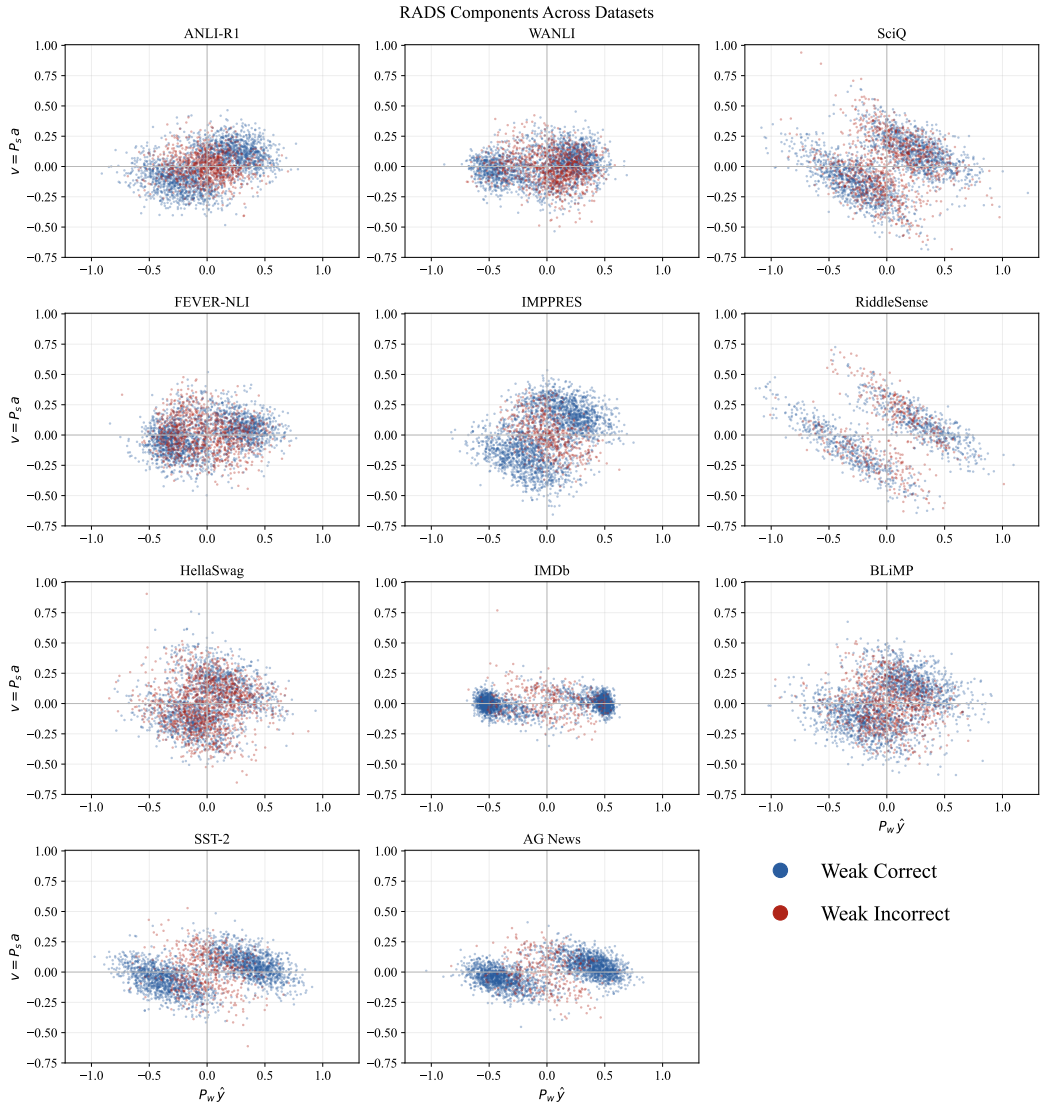


Figure 5: Additional scatters of $\mathbf{v}$ against $P_w\hat{\mathbf{y}}$ for more datasets. RiddleSense and SciQ exhibit a low PGR difference, while IMPPRES and FEVER-NLI exhibit high PGR difference.

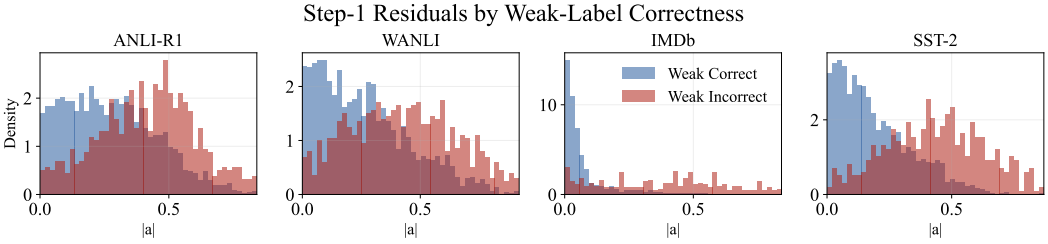


Figure 6: The value of $|a|$, coming from the output of step 1 in RADS, shows a separation between correctly and incorrectly labeled samples in a number of datasets.

Dataset	CKA	$\text{corr}(\hat{\mathbf{y}} - \mathbf{a}, \mathbf{v})$	Random	RADS high	RADS low
FEVER-NLI	0.94	+0.17	0.04 ± 0.05	0.32 ± 0.00	-0.69 ± 0.00
QNLI	0.93	+0.36	0.45 ± 0.05	0.44 ± 0.05	0.26 ± 0.04
IMPPRES	0.89	+0.30	0.35 ± 0.08	0.22 ± 0.03	-1.01 ± 0.04
HellaSwag	0.89	+0.11	0.44 ± 0.07	0.58 ± 0.01	0.14 ± 0.14
WANLI	0.89	+0.11	0.24 ± 0.04	0.62 ± 0.02	0.17 ± 0.02
TweetEval	0.87	+0.21	0.59 ± 0.16	0.65 ± 0.10	-0.33 ± 0.04
ANLI-R1	0.83	+0.37	0.51 ± 0.04	0.79 ± 0.01	0.20 ± 0.02
AG News	0.78	+0.30	0.41 ± 0.08	0.36 ± 0.08	-0.23 ± 0.19
SciNLI	0.78	+0.27	0.47 ± 0.09	0.42 ± 0.08	-0.39 ± 0.13
ETHICS justice	0.77	+0.17	0.24 ± 0.02	0.35 ± 0.02	0.03 ± 0.03
CosmosQA	0.77	+0.09	0.68 ± 0.06	0.75 ± 0.02	0.56 ± 0.06
ANLI-R3	0.76	+0.30	0.49 ± 0.06	0.57 ± 0.05	0.24 ± 0.07
SciTail	0.73	+0.32	0.46 ± 0.06	0.46 ± 0.03	0.18 ± 0.07
SST-2	0.66	+0.19	0.74 ± 0.05	0.79 ± 0.02	0.69 ± 0.01
BoolQ	0.61	+0.05	0.46 ± 0.09	0.53 ± 0.11	0.31 ± 0.04
QuAIL	0.51	+0.05	0.82 ± 0.01	0.80 ± 0.05	0.82 ± 0.00
IMDb	0.49	+0.05	0.77 ± 0.03	0.79 ± 0.02	0.72 ± 0.02
BLiMP	0.49	+0.17	-0.30 ± 0.07	-0.32 ± 0.12	-0.45 ± 0.05
CSQA	0.45	+0.04	0.92 ± 0.02	0.92 ± 0.00	0.84 ± 0.02
SciQ	0.45	+0.05	0.97 ± 0.01	0.97 ± 0.01	0.92 ± 0.01
AQuA-RAT	0.39	+0.04	0.59 ± 0.05	0.56 ± 0.03	0.51 ± 0.02
Wiki Toxic	0.36	+0.16	0.37 ± 0.10	0.12 ± 0.04	0.48 ± 0.02
RiddleSense	0.34	−0.15	0.80 ± 0.05	0.86 ± 0.04	0.78 ± 0.05

Table 7: A complete table of statistics regarding CKA for a variety of datasets and the associated PGR for RADS.

Dataset	RADS high	RADS low	RADS-Lin high	Random
ANLI-R1	0.67 ± 0.05	0.12 ± 0.04	0.78 ± 0.02	0.38 ± 0.04
WANLI	0.37 ± 0.15	-0.08 $\pm$ 0.04	0.16 ± 0.07	-0.13 $\pm$ 0.10
FEVER-NLI	0.11 ± 0.01	-0.34 $\pm$ 0.01	0.14 ± 0.07	-0.11 $\pm$ 0.02

Table 8: PGRs with Qwen2.5-1.5B as the weak model and Qwen2.5-7B as the strong model.

C.4 Relationships between $\mathbf{a}$ and $\mathbf{v}$

The scatters of $P_w\hat{\mathbf{y}}$ and $\mathbf{v}$ for other datasets tested can be seen in Figure 5. Consistently, we observe that the datasets with the highest PGR show a roughly unimodal distribution of samples, while those with the lowest PGR have some of the strongest bimodal distributions. This supports the idea that if the weak model labels confidently based on incorrect features, the strong model will not be able to improve the labels.

Additionally, we provide a complete table of the CKA for the datasets tested along with the attained PGRs when using RADS with samples where $|\mathbf{v}|$ is high and where $|\mathbf{v}|$ is low. The results are in Table 7. We also report the correlation between $\hat{\mathbf{y}} - \mathbf{a}$ and $\mathbf{v}$, connecting to the scatters seen in Figure 5.

Behavior of $\mathbf{a}$. We have identified that scoring points based on their relationships to the strong and weak model representations results in the greatest gains in performance. In the algorithm, we compute both $\mathbf{a}$ and $\mathbf{v}$, each capturing different properties of the data. In investigating the learned structure of both, we identified that $\mathbf{a}$ acts as a separator of incorrect and correct weak points. This is seen in Figure 6. For each of the 4 datasets shown, correct points have a lower value of $|\mathbf{a}|$ than incorrect points on average. This also matches the simple Gaussian synthetic setting used to produce Figure 1a. While these results show that we can use this method to filter data to have more correct points, the results in Figure 1b still hold: weak label accuracy is not a good predictor of performance.

	High band		Low band		
Dataset	wd 30	wd 0	wd 30	wd 0	Random
ANLI-R1	0.79 ± 0.01	0.62 ± 0.01	0.20 ± 0.02	0.32 ± 0.00	0.51 ± 0.04
WANLI	0.62 ± 0.02	0.28 ± 0.01	0.17 ± 0.02	0.29 ± 0.02	0.24 ± 0.04
FEVER-NLI	0.32 ± 0.00	0.22 ± 0.05	-0.69 ± 0.00	-0.18 ± 0.02	0.04 ± 0.05

Table 9: PGR Performance on RADS, comparing using weight decay to not using weight decay. We observe that weight decay is vital to performance.

	High band		Low band		
Dataset	wd 10	wd 0	wd 10	wd 0	Random
ANLI-R1	0.53 ± 0.02	0.64 ± 0.06	0.33 ± 0.07	0.23 ± 0.01	0.51 ± 0.04
WANLI	0.43 ± 0.04	-0.87 ± 0.01	-0.08 ± 0.17	-0.39 ± 0.12	0.24 ± 0.04
FEVER-NLI	0.33 ± 0.03	0.21 ± 0.09	-0.63 ± 0.04	-0.08 ± 0.03	0.04 ± 0.05

Table 10: PGR Performance on mlpstep2, comparing using weight decay to not using weight decay. We observe that weight decay is less important than for RADS yet still improves performance over no weight decay.

C.5 A Different Weak Labeler

The experiments in the main body focus on using the 0.5B and 7B Qwen2.5 models for the weak and strong models, respectively. We conduct some experiments using a different base model, the 1.5B Qwen2.5 model, to observe whether the gain with RADS holds for other model pairs.

The results are in Table 8. We observe that, for this model pair, RADS and RADS-Lin still outperform random sampling on these three datasets. In addition, RADS-Lin has a better PGR on two of the datasets. This is possibly explained by the larger model having a better linearly decodable representation than the smaller 0.5B model. If so, using a linear probe for step 1 is more appropriate.

C.6 The Importance of Weight Decay

Prior work with **PredGap** studied W2S from the perspective of ridge regression. In this setting, regularization plays a vital role in performance, both empirically and theoretically.

When establishing RADS with nonlinear alignment techniques, we maintained regularization in the form of weight decay. However, we have yet to establish that this is empirically a better choice. We experiment with using weight decay and removing weight decay for both RADS and mlpstep2.

The results can be found in Tables 9 and 10. In both cases, using weight decay improved performance noticeably. We predict this is due to overfitting, where the MLPs can almost always find a boundary in this high-dimensional space that nearly perfectly fits the labels. We need some mechanism to control this noise.

C.7 Data Scale Sweep for Another Dataset

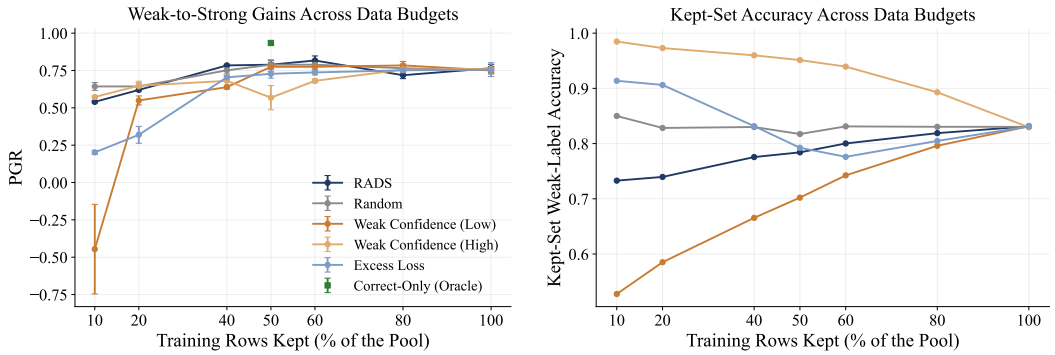


Figure 7: SST-2 against the fraction of the pool kept, five selectors and the correct-only oracle at 50%: PGR over three train seeds (left) and weak-label accuracy of the kept set (right).

In the main body, we used ANLI-R1 as a core dataset to compare different methods. This dataset was prototypical for representation alignment, showing the structure of $\hat{\mathbf{y}} - \mathbf{a}$ and $\mathbf{v}$ that is most common. We also tested this sweep on SST-2. In Table1, we observe that RADS is *not* the highest-performing method, and is instead CCA for a selection ratio of 50%.