Weak-to-strong Generalization via Formative Learning from Student Demonstrations & Teacher Evaluation

Anonymous Author(s)
Affiliation
Address
email

Abstract

As Large Language Models (LLMs) exceed human capabilities, providing reliable human feedback for evaluating and aligning them, via standard frameworks such as Reinforcement Learning from Human Feedback, becomes challenging. This raises a fundamental question: *how can we leverage weaker (teacher) supervision to elicit the full capabilities of a stronger (student) model?* This emerging paradigm, known as Weak-to-Strong (W2S) generalization, however, also introduces a key challenge as the strong student may "overfit" to the weak teacher's mistakes, resulting in a notable performance degradation compared to learning with ground-truth data. We show that this overfitting problem occurs because learning with weak supervision implicitly regularizes the strong student's policy toward the weak reference policy. Building on this insight, we propose a novel learning approach, called Weak Teacher **Ev**aluation of Strong Student D**e**monstrations or EVE, to instead regularize the strong student toward its reference policy. EVE's regularization intuitively elicits the strong student's knowledge through its own task demonstrations while relying on the weaker teacher to evaluate these demonstrations – an instance of formative learning. Extensive empirical evaluations demonstrate that EVE significantly outperforms existing W2S learning approaches and exhibits significantly better robustness under unreliable feedback compared to contrastive learning methods such as Direct Preference Optimization.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) [23, 5] has been a canonical framework for steering language models (LMs) to align with human values based on human demonstrations. This framework has demonstrated impressive performance across a wide range of tasks, from conversation to coding, where humans "can" provide reliable supervision. In the future, as these AI models reach or exceed human capabilities, they will be capable of solving complex tasks that are difficult for humans to supervise. For example, when these AI models acquire the ability to generate a code project with millions of lines of code or summarize an entire book with thousands of pages, humans are unlikely to provide reliable feedback to align these superhuman AI models effectively.

How can we align these superhuman AI models given the likely unreliable human supervision? Burns et al. [4] study this question by using a smaller LLM to represent unreliable human supervision on binary classification tasks. Effectively, this "weaker" teacher is prone to make mistakes when supervising a "stronger" student model. They observed a phenomenon called *weak-to-strong (W2S) generalization* – a stronger model finetuned with labels generated by a weaker model could outperform this weaker teacher without even seeing the ground truth labels. Despite the promising results, a key challenge in learning from weak supervision is the risk of overfitting [4], where the strong student inevitably learns to imitate the errors of the weak teacher. Burns et al. [4] study early-stopping as an

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

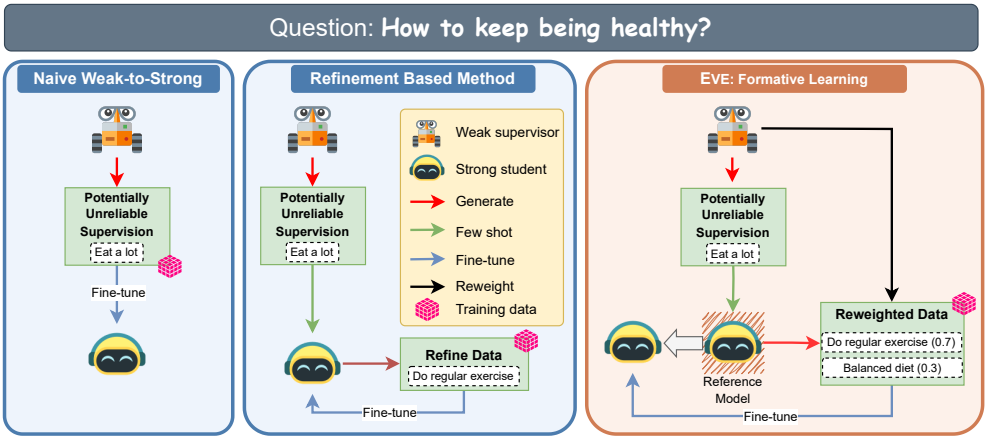

Figure 1: **EVE and existing W2S generalization methods**. Naive learning overfits the weak reference model, potentially imitating its mistakes (e.g., *"Eat a lot"*). Refinement learning "refines" the weak supervision (i.e., *"Do regular exercise"*). In contrast, EVE leverages the weak teacher as a reward function while eliciting the student's reference model salient knowledge

implicit regularization to prevent overfitting, but notes that early-stopping does not constitute a valid method as it unrealistically requires ground-truth labels.

This paper first provides a crucial theoretical insight into the overfitting problem in W2S generalization. Specifically, by representing the weak teacher as an Energy-Based Model (EBM), we reveal that learning from weak supervision involves maximizing the reward while simultaneously regularizing the strong student's policy toward the weak reference model. This process leads to a drawback: the strong student not only inherits the informative supervision but also amplifies the errors of the weak teacher, ultimately degrading the student's overall performance on the desired tasks [14].

Building upon this insight, we propose a novel learning method, called Weak Teacher **Ev**aluation of Strong Student D**e**monstrations (EVE), to enable the strong student to elicit its own (prior) knowledge of the task while relying on the weak teacher to evaluate, or score, such demonstrations – an instance of formative learning, effectively utilizing both the knowledge of the weak teacher and the student's reference model. As depicted in Fig. 1, EVE utilizes the weak teacher's demonstrations to prompt the strong student, allowing it to generate its own training data reflecting its understanding of the tasks. The generated samples are then adjusted by the logarithmic ratio of the weak teacher's policy pre- and post-alignment, which serves as a reward signal to guide the strong student's learning.

In summary, (1) we provide a theoretical characterization of overfitting in W2S learning; then (2) we introduce EVE, an approach that enables learning from strong student demonstrations, where the weak teacher acts as a reward function to evaluate the strong student's outputs; finally, (3) we show that EVE significantly outperforms naive W2S learning by overcoming the overfitting issue, demonstrating the effectiveness of utilizing the strong student's critical thinking ability under the weak teacher's reward evaluation; surprisingly, when learning from a weak and unreliable reward signal, EVE – an off-policy method – achieves significantly better performance and robustness to contrastive learning methods such as DPO [30].

2 Related Work

2.1 Weak-to-strong Generalization

Burns et al. [4] introduce a synthetic setup to study whether a stronger model can generalize well with weaker supervision, compared to training with high-quality or ground-truth data. Prior efforts investigate W2S phenomena only in binary classification setups, leaving other practical alignment-relevant tasks (e.g., open-ended text generation whose output has no fixed length and requires sharing vocabulary size between the strong student and weak teacher) largely under-explored [43, 6, 1]. Another line of work [34, 44, 45] leverages the pre-trained knowledge of the strong student to refine

labels curated from the weak teacher, thereby improving the supervision quality. Ye et al. [44] study W2S generalization on text-generation tasks, where they simulate *unreliable demonstrations* and *unreliable comparison feedback* during the alignment phase.

Different from the prior work, this paper extends W2S generalization beyond classification. We elicit the latent knowledge of the strong student about the intended tasks, which is then evaluated by the weak teacher's reward model. Additionally, by interpreting learning from weak supervision as reward maximization, our approach generalizes refinement-based methods [44, 41].

‘

2.2 Reinforcement Learning from Human Feedback

RLHF aims to align LMs with human preferences and values [5, 3], and has demonstrated impressive performance on established benchmarks [22, 15, 39, 40, 38]. However, the RLHF pipeline incurs significant computational costs and requires a large amount of high-quality human preference labels.

Recent advancements, such as Direct Alignment Algorithms (DAAs) [30, 37], bypass the need for an explicit reward model and directly train the LMs on the human preference data. Reinforcement Learning with AI Feedback [25] uses a well-trained language model (e.g., GPT-4 or Claude-3.5 Sonnet) to provide preference feedback as a substitute for human supervision. More recently, Ye et al. [44] study whether standard RLHF remains effective under unreliable feedback.

We demonstrate that **contrastive-learning** approaches [31, 2] heavily suffer from the reward over-optimization issue [31, 10]. In contrast, EVE – also an offline supervised approach – is significantly more robust to unreliable feedback and achieves a better reward-KL tradeoff than DAAs. This finding is significant as it contradicts observations in prior work [36], which shows that DAAs with negative gradient perform significantly better than offline supervised methods in conventional alignment scenarios with human feedback.

2.3 Reward Maximization with KL Regularization as Distributional Matching

Prior works show that reward maximization with KL regularization in standard RLHF can be viewed as minimizing the reverse KL between the LM policy π_θ and the target distribution that represents the aligned language model [17, 16, 12]. Other studies also explored the use of forward KL, which corresponds to setting the reward maximization as supervised learning [21, 28]. Similarly, our paper shows that imitating a weak teacher can be viewed as reward maximization, where the reward is defined as the log probability of the weak teacher, with a KL regularization toward the weak reference model, causing the over-optimization problem.

3 Preliminaries

3.1 LLM Alignment with Human Preferences

LLM alignment can be viewed as reward-maximization with KL-constrained:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[r(x,y)\right] - \beta \mathrm{KL}(\pi_\theta || \pi_{\text{ref}}) \tag{1}$$

where y is a sampled response from π_θ, β controls the trade-off between maximizing the reward and deviation from the reference model π_{ref}, and r is the reward function that captures human preferences.

3.2 Offline Fine-Tuning Methods for Reward Maximization

Directly optimizing the objective in Eq. (1) requires repeated sampling, which is computationally expensive. Alternatively, equivalent offline methods fall into 2 main categories:

Contrastive Learning Methods. Approaches, such as DPO [30] and IPO [2], directly update the LM policy π_θ on human preference data. These methods represent the reward implicitly via the LM π_θ and the reference model π_{ref} as:

$$r_\theta(x,y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \tag{2}$$

where $Z(x) = \sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r_\theta(x, y)\right)$ is the normalization factor. Using this representation, a general objective can be derived to train the policy on human preference data, as follows:

$$\mathcal{L}(\pi_\theta, \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[f\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{3}$$

where f is a convex loss function. The gradient of contrastive learning approaches, therefore, consists of a **positive gradient** term that increases the likelihood of the preferred response y_w and a **negative gradient** term that pushes down the likelihood of the non-preferred response y_l.

Offline Supervised Methods. This alternative class of methods, including RAFT [8] and RWR [28], minimizes a weighted maximum likelihood objective. Formally, these methods first sample K completions per prompt x from the reference model π_{ref}, i.e., $y_{1,\cdots,K} \sim \pi_{\text{ref}}(\cdot|x^{(i)})$. These responses are then weighted by a non-negative weighting function $F\left(x, y_k|y_{1,\cdots,K}\right)$ conditioned on the other sampled responses and maximize:

$$\max_{\pi_\theta} \mathbb{E}_{(x,y_{1,\cdots K})\sim\mathcal{D},\pi_{\text{ref}}(\cdot|x)}\left[\sum_{k=1}^{K} \log \pi_\theta(y_k|x) \cdot F\left(x, y_k|y_{1,\cdots,K}\right)\right]$$

Intuitively, since $F\left(x, y|y_{1,\cdots K}\right)$ is always non-negative, these methods always increase the likelihood of responses generated from π_{ref}. Responses that are more preferred will be assigned higher weights, there is no **negative gradient** effect to push down the likelihood of suboptimal responses.

3.3 Weak-to-Strong Evaluation Pipeline

We review the W2S evaluation pipeline in [4], which consists of three stages, as follows:

(1) Weak Teacher Creation: The weak teacher is created by fine-tuning a small pre-trained model to align with human preferences. We utilize SFT+DPO, a standard preference learning pipeline, to ensure the weak model acquires knowledge about alignment tasks. The resulting model is denoted as π^{weak}.

(2) Strong Student Learning with Weak Supervision: The weak model is then used to generate weak supervision data $\mathcal{D}_{\text{weak}} = \{x^{(i)}, y^{(i)}\}$ where $x^{(i)}$ and $y^{(i)}$ are the prompt and the generated response from π^{weak}, respectively. The strong model π_θ is then fine-tuned using the weak supervision data with the SFT objective.

(3) Strong Student Learning with Ground-truth Supervision: Another strong model π^{strong} is fine-tuned with the Ground-truth human labels to establish the upper-bound performance. To ensure that this aligned model fully acquires the target task's capabilities, it goes through an additional, preference learning phase (e.g., DPO).

The W2S generalization performance of π_θ can be measured by Performance Gap Recovered (**PGR**):

$$\text{PGR} = \frac{\mathcal{P}_{\text{weak-to-strong}} - \mathcal{P}_{\text{weak}}}{\mathcal{P}_{\text{strong}} - \mathcal{P}_{\text{weak}}}$$

where $\mathcal{P}_{\text{weak-to-strong}}$, $\mathcal{P}_{\text{weak}}$, and $\mathcal{P}_{\text{strong}}$ are the task performance of π_θ, π^{weak}, and π^{strong}, respectively.

4 Formative Learning with EVE

4.1 Learning from Weak Supervision Implicitly Aligns with Weak Reference Model

This section connects W2S learning to reward maximization and builds the theory behind the model's behavior, i.e., its generalization characteristics.

We begin by representing the weak teacher in the form of energy-based models [30, 18, 13]:

$$\pi^{\text{weak}}(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}^{\text{weak}}(y|x) \exp\left(r^{\text{weak}}(x, y)/\beta\right)$$

where $\pi_{\text{ref}}^{\text{weak}}$ is the SFT version of π^{weak}.

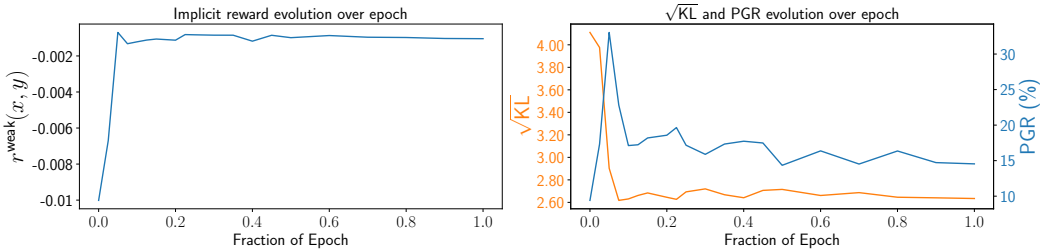

Figure 2: **Learning from weak supervision** as reward maximization. **Left:** the strong model π_θ learns to maximize the implicit reward $r^{\text{weak}}(x, y) = \beta \log \pi^{\text{weak}}_{\text{align}}(y|x)$ - $\beta \log \pi^{\text{weak}}_{\text{ref}}(y|x)$. **Right:** the strong model also learns to imitate the weak reference model $\pi^{\text{weak}}_{\text{ref}}$'s mistakes, leading to performance degradation (in PGR).

Proposition 4.1. *W2s generalization with a weak teacher $\pi^{weak}(y|x)$ and a strong student π_θ (the training model) can be cast as the following optimization problem:*

$$\min_{\pi_\theta} KL\left(\pi^{weak}||\pi_\theta\right) \tag{4}$$
$$s.t\ \pi^{weak} = \arg\min_{\pi} KL\left(\pi||\pi^{EBM}\right)$$

where $\pi^{EBM}(y|x) \propto \pi^{weak}_{ref}(y|x) \exp\left(r(x, y)/\beta\right)$.

The proof is straightforward and deferred to the Appendix D.2. This shows that imitating the weak teacher can be seen as finding an EBM policy π^{EBM}, which is the optimal solution in the lower-level objective. This leads to the following theorem.

Theorem 4.2. *The optimal solution to W2S generalization is equivalent to the optimal solution in the following objective:*

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[r^{weak}(x, y)\right] - \lambda KL(\pi_\theta||\pi^{weak}_{ref}) \tag{5}$$

Proof Sketch. Notice that the objective for training the strong student, and the reverse KL share the same optimal solution π_θ. In addition, it can be shown that minimizing the reverse KL between the strong student and the weak teacher,

$$\min_{\pi_\theta} \text{KL}\left(\pi_\theta||\pi^{\text{weak}}\right), \tag{6}$$

is equivalent to maximizing the KL-constrained reward objective in Eq. (5). □

Theorem 4.2 provides a key insight: imitating the weak teacher maximize an implicit reward, $r^{\text{weak}}(x, y) = \beta \log \pi^{\text{weak}}(y|x) - \beta \log \pi^{\text{weak}}_{\text{ref}}(y|x)$, while regularizing (with KL objective) the strong student toward the weak reference model $\pi^{\text{weak}}_{\text{ref}}$. Consequently, instead of aiming to elicit knowledge of the strong student, existing W2S learning remains confined to the knowledge of the weak model, which may adversely impact the strong student's performance.

4.2 Suboptimal Weak-to-Strong Generalization toward Weak Reference Model

We empirically confirm the theoretical insight in the previous section. Specifically, we analyze the W2S training progression on $\mathcal{D}_{\text{weak}}$: at each checkpoint, we generate responses using the corresponding intermediate model with the same set of prompts, from which we calculate the implicit reward $r^{\text{weak}}(x, y) = \beta \log \pi^{\text{weak}}(y|x) - \beta \log \pi^{\text{weak}}_{\text{ref}}(y|x)$, the divergence $\text{KL}(\pi_\theta||\pi^{\text{weak}}_{\text{ref}})$, and the PGR.

Fig. 2 shows that while the strong model learns to maximize the implicit reward (Left), the learned policy is also regularized towards the weak reference model $\pi^{\text{weak}}_{\text{ref}}$, indicated by the consistently low KL divergence $\text{KL}(\pi_\theta||\pi^{\text{weak}}_{\text{ref}})$ shortly after the training progresses (Right). Moreover, we also observe that the PGR, as measured by the golden reward function, decreases significantly (Right). This suggests that imitating the weak reference model $\pi^{\text{weak}}_{\text{ref}}$ (and potentially inheriting its mistakes) negatively impacts the performance of the strong student.

4.3 EVE: Eliciting Strong Student Knowledge

Motivated by the connection between imitating the weak teacher and reward maximization, we "generalize" the KL-constrained reward maximization learning of the strong student π:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[r^{\text{weak}}(x,y)\right] - \lambda \text{KL}(\pi_\theta || \hat{\pi}) \tag{7}$$

where λ controls the trade-off between maximizing the reward and deviation from a regularization policy $\hat{\pi}(y|x)$. Next, we propose one specific choice of the regularization policy $\hat{\pi}$ that can facilitate the elicitation of the strong student's knowledge, thereby enhancing W2S generalization.

The choice of regularization policy $\hat{\pi}$. Burns et al. [4] interpret W2S generalization in terms of saliency: some tasks are already salient to the strong student; in this view, the role of the weak teacher is to elicit the student's latent knowledge rather than enforcing naive imitation of the weak teacher's own demonstrations. Inspired by this interpretation, we propose to regularize the learning policy toward the strong student pre-trained model, i.e., $\hat{\pi}(y|x) = \pi_{\text{ref}}^{\text{strong}}(y|x)$. This design choice serves an important goal: to encourage the learned policy π_θ to remain close to the initial strong reference model $\pi_{\text{ref}}^{\text{strong}}$, thereby facilitating the elicitation of the student's prior knowledge while simultaneously incorporating assessment from the weak teacher. Similar to [4], to elicit the strong student's knowledge of the task, we first create the weak teacher's demonstrations, which are then used in few-shot prompting the strong reference model $\pi_{\text{ref}}^{\text{strong}}$ to generate task-relevant outputs, as $\pi_{\text{ref}}^{\text{strong}}$ is not trained to follow instructions. We provide detailed examples in Appendix B.6.

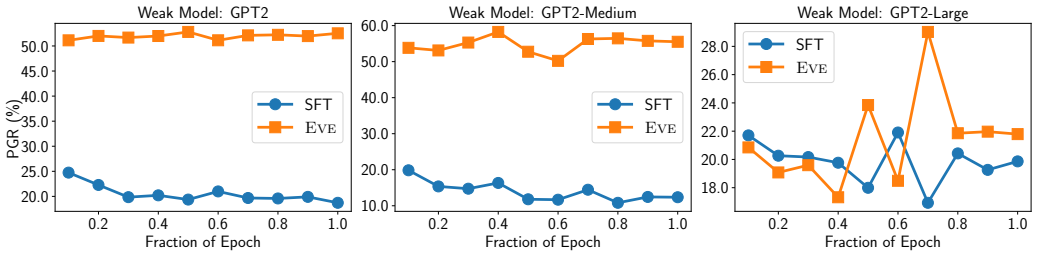

Figure 3: Evolution of PGR (%). We observe clear signs of overfitting to the weak teacher's errors well before finishing a single epoch. Notably, when there is a large gap between the strong student and the weak teacher, the student reaches its best performance within the first 10% of the epoch. EVE has little to no PGR degradation and significantly outperforms naive W2S learning (SFT).

Optimization. Directly optimizing the objective in Eq. (7) can incur significant computational costs, as it requires repeated sampling from the strong student π_θ inside the training loop [30]. Following prior work [30, 28, 27], it is straightforward to show that the optimal policy to this KL-constrained objective takes the form:

$$\pi_r(y|x) = \frac{1}{Z(x)} \exp\left(r(x,y)/\lambda\right) \pi_{\text{ref}}^{\text{strong}}(y|x)$$

where $Z(x) = \sum_y \pi_{\text{ref}}^{\text{strong}}(y|x) \exp\left(r(x,y)/\lambda\right)$ is the normalization constant. We can also leverage the duality between the reward function and the weak teacher π^{weak} [30]. Given the optimal policy π_r, we can then formulate a supervised learning objective for the parametrized strong student π_θ to match with this optimal policy, resulting in the following objective:

$$\max_{\pi_\theta} \mathcal{J}(\pi_\theta) = \max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_{\text{ref}}^{\text{strong}}(\cdot|x)} \left[\frac{\left(\pi^{\text{weak}}(y|x)/\pi_{\text{ref}}^{\text{weak}}(y|x)\right)^{\beta/\lambda}}{Z(x)} \cdot \log \pi_\theta(y|x)\right]$$

where the β/λ ratio controls the impact of the weak-supervision reward signal during the strong student's updates. A high β/λ ratio leads to a more uniform update, where all samples are assigned similar weights; i.e., there will be no weak supervision in learning. Conversely, a low β/λ ratio results in a more focused policy update that prioritizes samples with high weak-supervision reward signals. This objective avoids sampling directly from π_θ on every update as π_θ changes during

training; instead, we can sample the responses from the fixed $\pi_{\text{ref}}^{\text{strong}}$ once at the beginning of the optimization, which is significantly more efficient.

We also estimate the intractable normalization factor $Z(x)$ using *Self-Normalizing Importance Sampling* [24]. Formally, given $K > 1$ i.i.d. completions $y^1, \cdots, y^N \sim \pi_{\text{ref}}^{\text{strong}}(\cdot|x)$ drawn from strong reference model, we can define an empirical distribution by normalizing the log-ratio $f(x, y) = \frac{\beta}{\lambda}\left(\log \pi^{\text{weak}}(y|x) - \log \pi_{\text{ref}}^{\text{weak}}(y|x)\right)$ over K samples:

$$F(x, y^i|y^{1,\cdots,K}) = \frac{K \cdot \exp\left(f(x, y^i)\right)}{\sum_{k=1}^{K} \exp\left(f(x, y^k)\right)} \tag{8}$$

where the normalization is estimated by $Z(x) \approx \frac{1}{K}\sum_{k=1}^{K} \exp\left(f(x, y^k)\right)$. In summary, the final estimate is:

$$\mathcal{J}(\pi_\theta) = \mathbb{E}_{x\sim\mathcal{D}, y^{1,\cdots,K}\sim\pi_{\text{ref}}^{\text{strong}}(\cdot|x)}\left[\log \pi_\theta(y^i|x) \cdot\ F(x, y^i|y^{1,\cdots,K})\right]$$

We refer to this W2S learning approach as EVE. EVE can be seen as an offline supervised method, where the weighting function is the exponential of the implicit reward defined in Eq. (2).

5 Experiments

In this section, we empirically evaluate EVE's W2S generalization performance on two tasks: **controlled-summarization** and **instruction following**.

5.1 Controlled-Summarization

Setup. We choose the representative Reddit TL;DR summarization [35] dataset and follow the synthetic setup from [10, 47, 30], where we train a *golden* reward model $r_{\text{gold}}(x, y)$ to label synthetic preference data $\mathcal{D}_{\text{golden}}$ for fine-tune weak-aligned model and evaluation. We use GPT2-series [29] (GPT2-Base/Medium/Large) as weak teachers and a more advanced LLama-3.2-3B model [19, 20] as the strong student. The weak model π^{weak} is the aligned model with DPO [30] from $\mathcal{D}_{\text{golden}}$.

Baselines. In additional to EVE, we evaluate several existing W2S approaches, including **SFT** – which naively fine-tunes the strong student on weak supervision data $\mathcal{D}_{\text{weak}}$ – and (2) **Refinement** [34, 41] – which prompts the strong student to refine the responses generated by the weak teacher and fine-tunes the strong student with the refined responses.

Results. Fig. 4 shows the PGR results. EVE consistently outperforms the other baselines across all weak teachers. Notably, under the supervision of GPT-2 (the weakest model), EVE achieves a nearly 25% performance boost over SFT. Moreover, SFT achieves the peak performance early in training (around 10% of the epoch), but its performance steadily declines thereafter. In contrast, **EVE demonstrates minimal to no degradation in PGR over the course of the training process**. As discussed in Section 4, this can be attributed to the ability of EVE to more effectively balance learning from the weak teacher and the salient knowledge of the strong reference model.

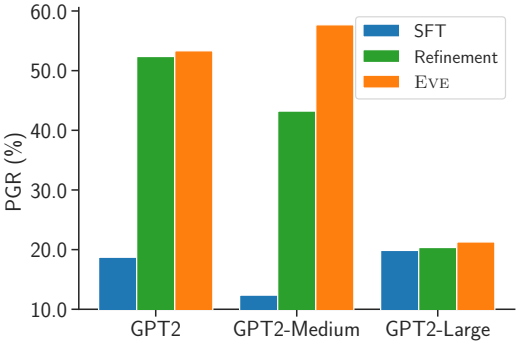

Figure 4: PGR (%) of SFT, Refinement and EVE.

Impact of β/λ ratio. We investigate the impact of β/λ on W2S performance. Fig. 5 illustrates the impact of β/λ on PGR across different weak teachers. Setting β/λ around 1.0 achieves optimal or near-optimal performance. Consequently, we default $\beta/\lambda = 1.0$ in all experiments, **eliminating the need for hyperparameter tuning that requires ground-truth labels**. Without the weak supervision (i.e., $\beta/\lambda = \infty$), the performance significantly decreases; this confirms the benefit of learning from the weak teacher's reward signals. Conversely, setting β/λ to a very low value can

also degrade the performance. One possible explanation is that, as $\beta/\lambda \to 0$, the weighting function $F(x, y^i | y^{1,\cdots,K})$ converges to a one-hot distribution, where the response with the highest reward is assigned a weight of 1 and the rest are ignored. This limits learning from a few samples, making it susceptible to simply memorizing the training data [26].

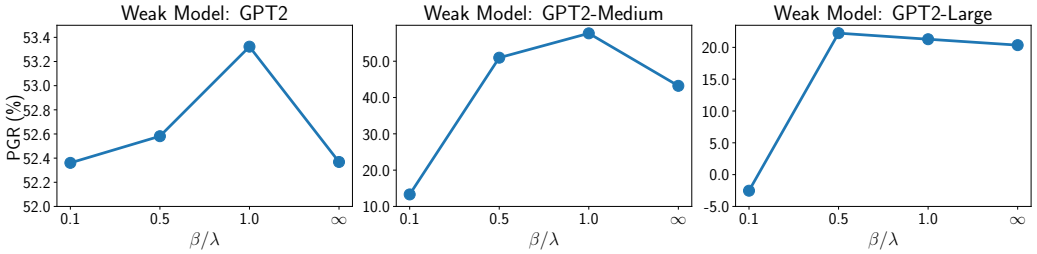

Figure 5: PGR (%) of various β/λ ratios in EVE's objective.

Scaling dataset size. We additionally study the impact of scaling the number of responses K per prompt. Fig. 6 shows the performance of EVE and SFT. EVE demonstrates improved performance as we increase the size of the training dataset (especially as the weak teacher is stronger), while SFT's performance decreases. This can be explained by the fact that as the training data size increases, the strong student also becomes more susceptible to learning the weak teacher's mistakes. In contrast, EVE is designed to avoid this overfitting problem, thus, it can leverage the increased supervision significantly better.

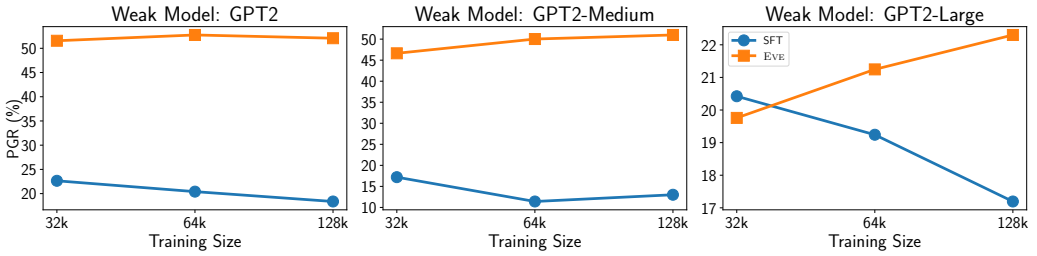

Figure 6: Scaling the training size (32k, 64k and 128k) in EVE and SFT (trained for one epoch). EVE shows notable improvement as the training size increases, while SFT suffers from overfitting.

5.2 Instruction Following

Setup. We use Qwen2.5-7B as the strong student and Llama-3.2-1B as the weak teacher. The strong reference model $\pi_{\text{base}}^{\text{strong}}$ is initialized from the pre-trained distribution, and the weak model π^{weak} is fine-tuned with DPO on the UltraFeedback dataset [7].

Evaluation. We evaluate EVE on two standard instruction-following benchmarks, AlpacaEval 2.0 [9] and IFEval [46]. For AlpacaEval 2.0, we report length-controlled win-rates against `gpt4-turbo`, with `gpt-4o-mini` serving as the judge.

Baselines. We evaluate EVE against **SFT, Refinement** and **DPO** - which uses the weak teacher as reward signal to label preference data generated by the strong student. Following prior works [31, 11], we train DPO for 1 epoch with $\beta = 0.05$ as default hyperparameters.

Results. We report the results in Fig. 7. EVE consistently outperforms the other W2S approaches across all benchmarks. Interestingly, we find that weak supervision can provide a reliable signal for guiding the strong student, not only in encouraging instruction-following behavior but also helping to filter out non-compliant responses.

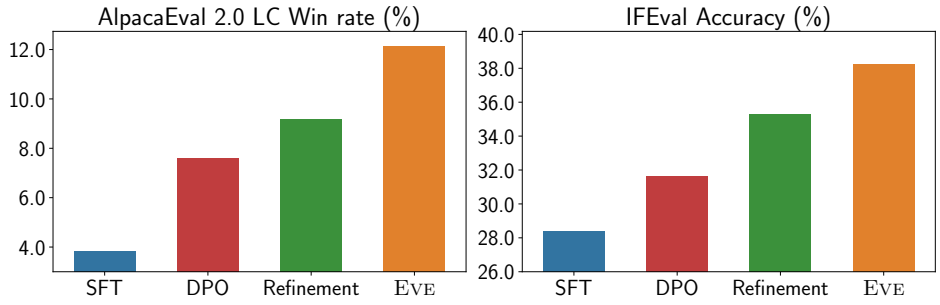

Figure 7: Results on various instruction following benchmarks for various methods.

5.3 MLE and Contrastive Learning in W2S Generalization

Fig. 7 also shows the advantages of EVE over contrastive learning approaches (e.g., DPO). Standard RLHF frameworks (e.g., PPO [33] and DPO [30]) can be seen as optimizing the reverse KL, while EVE optimizes the forward KL. As noted in [36], the reverse KL can modify the probability mass more aggressively than forward KL, resulting in a large deviation from π_{ref} to find the peak reward region. Conversely, the forward KL tends to deviate less from its initial distribution towards the peak reward. This might be beneficial in W2S learning due to the following reasons:

(i) Unreliable learning signal. Unlike standard RLHF, W2S's feedback is highly unreliable. Finding the peak reward region with the reverse KL can result in performance degradation due to over-optimization as observed in [44] (and re-confirmed in Fig 9 in our Appendix). Importantly, over-optimization can be more severe in W2S generalization as even humans cannot providea reliable signal to avoid these undesirable behaviors [4].

(ii) Already capable strong student. Similar to prior works [4, 44, 6], we assume that the strong student is already capable of solving the target tasks. Consequently, we hypothesize that the response region that achieved high rewards should be near the strong student. Therefore, the forward KL, inducing less deviation from the initial distribution, can be seen as an additional implicit regularization and performs better.

6 Limitations

We did not experiment with larger language models (> 7B) due to limited computational resources. Given the resource demands of generating data from the strong student, future work will focus on using the strong student for evaluation to eliminate the need for generating strong student data. For example, one direction is to explore extensions to our strategy proposed in Section D.1, which relies on the strong student only for reward evaluation. Tajwar et al. [36]. While our method does introduce additional memory overhead from teacher feedback calculations, this cost is relatively minimal compared to the overall training process of the strong student.

7 Conclusion and Discussion

This paper studies the W2S generalization and provides a new theoretical perspective on imitating the weak teacher. We show that imitating the weak teacher is equivalent to maximizing an implicit reward and regularizing the student towards the weak reference policy, which can amplify the bias or mistakes of this supervised fine-tuned weak teacher while not effectively eliciting knowledge from the strong student. Building upon this observation, we propose EVE, which directly optimizes the strong student using an RLHF objective with the "forward KL" regularization towards its latent knowledge of the given task. Extensive empirical results demonstrate that EVE achieves superior performance to existing W2S baselines and effectively mitigates the overfitting problem in W2S generalization.

References

[1] A. Agrawal, M. Ding, Z. Che, C. Deng, A. Satheesh, J. Langford, and F. Huang. Ensemw2s: Can an ensemble of llms be leveraged to obtain a stronger llm?, 2024. URL https://arxiv.org/abs/2410.04571.

[2] M. G. Azar, Z. D. Guo, B. Piot, R. Munos, M. Rowland, M. Valko, and D. Calandriello. A general theoretical paradigm to understand learning from human preferences. In *AISTATS*, pages 4447–4455, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

[3] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

[4] C. Burns, P. Izmailov, J. H. Kirchner, B. Baker, L. Gao, L. Aschenbrenner, Y. Chen, A. Ecoffet, M. Joglekar, J. Leike, I. Sutskever, and J. Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4971–5012. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/burns24b.html.

[5] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

[6] Z. Cui, Z. Zhang, W. Wu, G. Sun, and C. Zhang. Bayesian weaks-to-strong from text classification to generation, 2024. URL https://arxiv.org/abs/2406.03199.

[7] N. Ding, Y. Chen, B. Xu, Y. Qin, S. Hu, Z. Liu, M. Sun, and B. Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.183. URL https://aclanthology.org/2023.emnlp-main.183/.

[8] H. Dong, W. Xiong, D. Goyal, Y. Zhang, W. Chow, R. Pan, S. Diao, J. Zhang, K. SHUM, and T. Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=m7p5O7zblY.

[9] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2025. URL https://arxiv.org/abs/2404.04475.

[10] L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

[11] Z. Gao, J. D. Chang, W. Zhan, O. Oertell, G. Swamy, K. Brantley, T. Joachims, J. A. Bagnell, J. D. Lee, and W. Sun. REBEL: Reinforcement learning via regressing relative rewards. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024. URL https://openreview.net/forum?id=4SKidIUPP6.

[12] D. Go, T. Korbak, G. Kruszewski, J. Rozen, N. Ryu, and M. Dymetman. Aligning language models with preferences through f-divergence minimization. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11546–11583. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/go23a.html.

[13] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1352–1361. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/haarnoja17a.html.

[14] J. Hong, N. Lee, and J. Thorne. ORPO: Monolithic preference optimization without reference model. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.626. URL https://aclanthology.org/2024.emnlp-main.626/.

[15] e. a. Hugo Touvron, Louis Martin. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

[16] T. Korbak, H. Elsahar, G. Kruszewski, and M. Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=XvI6h-s4un.

[17] T. Korbak, E. Perez, and C. Buckley. RL with KL penalties is better viewed as Bayesian inference. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1083–1091, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.77. URL https://aclanthology.org/2022.findings-emnlp.77.

[18] S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review, 2018. URL https://arxiv.org/abs/1805.00909.

[19] MetaAI. Introducing llama 3.1: Our most capable models to date. 2024a. URL https://ai.meta.com/blog/meta-llama-3-1/.

[20] MetaAI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. 2024b. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

[21] M. Norouzi, S. Bengio, z. Chen, N. Jaitly, M. Schuster, Y. Wu, and D. Schuurmans. Reward augmented maximum likelihood for neural structured prediction. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/2f885d0fbe2e131bfc9d98363e55d1d4-Paper.pdf.

[22] OpenAI, J. Achiam, and e. a. Steven Adler. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[23] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[24] A. B. Owen. *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

[25] J.-C. Pang, P. Wang, K. Li, X.-H. Chen, J. Xu, Z. Zhang, and Y. Yu. Language model self-improvement by reinforcement learning contemplation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=38E4yUbrgr.

[26] S. Park, K. Frans, S. Levine, and A. Kumar. Is value learning really the main bottleneck in offline RL? In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024. URL https://openreview.net/forum?id=Rbf1h7NH11.

[27] X. B. Peng, A. Kumar, G. Zhang, and S. Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019. URL https://arxiv.org/abs/1910.00177.

[28] J. Peters and S. Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 745–750, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937933. doi: 10.1145/1273496.1273590. URL https://doi.org/10.1145/1273496.1273590.

[29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[30] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

[31] R. Rafailov, Y. Chittepu, R. Park, H. Sikchi, J. Hejna, W. B. Knox, C. Finn, and S. Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pf4OuJyn4Q.

[32] J. Schulman. Approximating kl divergence, 2020. URL http://joschu.net/blog/kl-approx.html.

[33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

[34] S. Somerstep, F. M. Polo, M. Banerjee, Y. Ritov, M. Yurochkin, and Y. Sun. A transfer learning framework for weak-to-strong generalization, 2024. URL https://arxiv.org/abs/2405.16236.

[35] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

[36] F. Tajwar, A. Singh, A. Sharma, R. Rafailov, J. Schneider, T. Xie, S. Ermon, C. Finn, and A. Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=bWNPx6t0sF.

[37] Y. Tang, Z. D. Guo, Z. Zheng, D. Calandriello, R. Munos, M. Rowland, P. H. Richemond, M. Valko, B. Avila Pires, and B. Piot. Generalized preference optimization: A unified approach to offline alignment. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 47725–47742. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/tang24b.html.

[38] Z. Wang, L. Hou, T. Lu, Y. Wu, Y. Li, H. Yu, and H. Ji. Enable lanuguage models to implicitly learn self-improvement from data. In *Proc. The Twelfth International Conference on Learning Representations (ICLR2024)*, 2024.

[39] W. Xiong, H. Dong, C. Ye, Z. Wang, H. Zhong, H. Ji, N. Jiang, and T. Zhang. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. In *Proc. The Forty-first International Conference on Machine Learning (ICML2024)*, 2024.

[40] W. Xiong, H. Dong, C. Ye, Z. Wang, H. Zhong, H. Ji, N. Jiang, and T. Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Proc. ICLR2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

[41] Y. Yang, Y. Ma, and P. Liu. Weak-to-strong reasoning. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8350–8367, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.490. URL https://aclanthology.org/2024.findings-emnlp.490/.

[42] Y. Yang, Y. Ma, and P. Liu. Weak-to-strong reasoning. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8350–8367, Miami, Florida, USA, Nov. 2024. Association for Computational

Linguistics. doi: 10.18653/v1/2024.findings-emnlp.490. URL https://aclanthology.org/2024.findings-emnlp.490.

[43] R. Ye, Y. Xiao, and B. Hui. Weak-to-strong generalization beyond accuracy: a pilot study in safety, toxicity, and legal reasoning, 2024. URL https://arxiv.org/abs/2410.12621.

[44] Y. Ye, C. Laidlaw, and J. Steinhardt. Iterative label refinement matters more than preference optimization under weak supervision. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=q5EZ7gKcnW.

[45] C. Zheng, Z. Wang, H. Ji, M. Huang, and N. Peng. Weak-to-strong extrapolation expedites alignment. In *arxiv*, 2024.

[46] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

[47] Z. Zhou, Z. Liu, J. Liu, Z. Dong, C. Yang, and Y. Qiao. 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*, 2024. URL https://openreview.net/forum?id=dOJ6CqWDf1.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper faithfully adheres to the claims and motivation in the abstract and provides proof and detailed empirical studies in support.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A discussion of our limitations can be found at section A Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide proofs and empirical evidence to support all our theoretical results

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed guidance on reproducibility by specifying all datasets and hyperparameters used in this work. Furthermore, we will release our GitHub implementation if the paper is accepted.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We've only used open-source models and open-source datasets for all experiments in our work. We provide details experiments in section B Appendix and Section 5 to reproduce the results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We list experiment details about the training and results details in Section 5 and Section B Appendix. Our experiments only use open-source datasets and models.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Training large language models is computationally expensive. Due to constraints in computational resources and financial budget, our experiments do not include multiple random seeds for each configuration. However, our evaluation protocol aligns with that of prior works in W2S generalization and RLHF. [44, 42, 31].

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information on computational resources in the experimental details section B in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss societal impacts in Section A of the appendix. We do no expect any negative societal impacts directly resulting from the contributions of our paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for the responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We use public models that are fine-tuned for alignment on open-source datasets. Our work does not contribute any risk to these public models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:We have properly cited papers and resources used in our experiment. The pretrained model comes from Llama-3.2-3B, which are classified under the Community License agreement: https://huggingface.co/meta-llama/Llama-3.2-3B. Both the TL:DR and UltraFeedback datasets used in this work use the MIT License.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We don't have experiments involving crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We use LLM for grammar checking only.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.