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

Phuc Nguyen [1] **Chinh Duc La** [1] **Heng Ji** [2] **Khoa D Doan** [1]

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 naive SFT and refinement approaches.

1. Introduction

Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Christiano et al., 2017) 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. (2024) 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 (Burns et al., 2024), where the strong student inevitably learns to imitate the errors of the weak teacher. Burns et al. (2024) study early-stopping as an 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 (Hong et al., 2024).

Building upon this insight, we propose a novel learning

[1]VinUniversity [2]University of Illinois Urbana-Champaign. Correspondence to: Khoa D Doan <khoa.dd@vinuni.edu.vn>.

Published at Data in Generative Models Workshop: The Bad, the Ugly, and the Greats (DIG-BUGS) at ICML 2025, Vancouver, Canada.

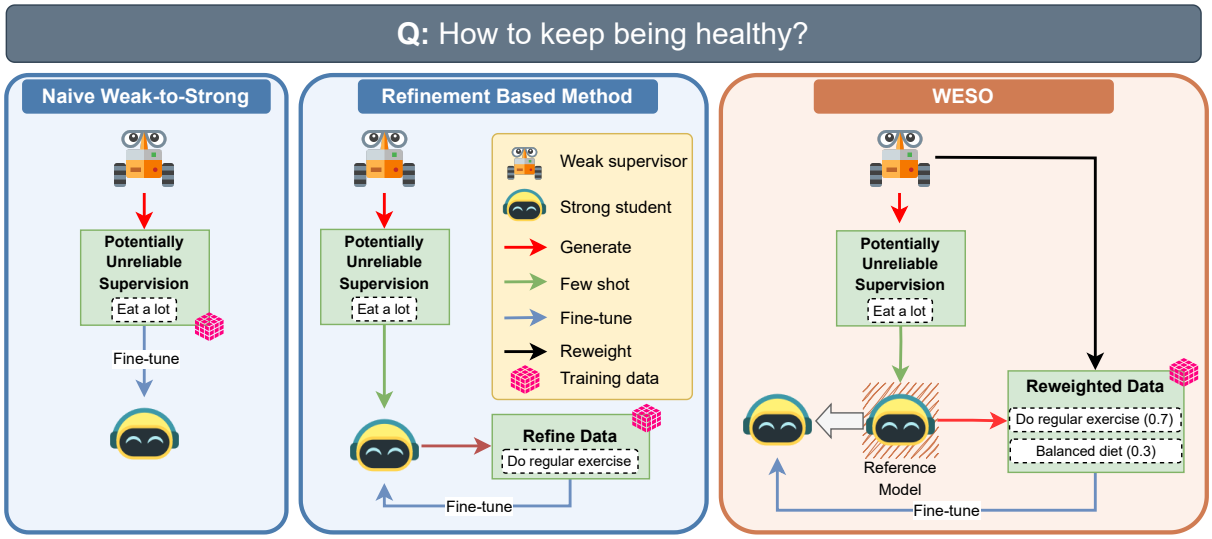

Figure 1: **EVE and existing W2S generalization methods**. Naive learning overfits the weak reference model, potentially imitating its mistakes (e.g., *"Eat a lot"*). Refinement learning "refines" the weak supervision (i.e., *"Do regular exercise"*). In contrast, EVE leverages the weak teacher as a reward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

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 to naive SFT and refinement approaches.

2. Related Work

2.1. Weak-to-strong Generalization

Burns et al. (2024) 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 (Ye et al., 2024; Cui et al., 2024; Agrawal et al., 2024). Another line of work (Somerstep et al., 2024; Ye et al., 2025; Zheng et al., 2024) 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. (2025) 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 (Ye et al., 2025; Yang et al., 2024).

2.2. Reinforcement Learning from Human Feedback

RLHF aims to align LMs with human preferences and values (Christiano et al., 2017; Bai et al., 2022), and has demonstrated impressive performance on established benchmarks (OpenAI et al., 2024; Hugo Touvron, 2023; Xiong et al., 2024a;b; Wang et al., 2024). 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) (Rafailov et al., 2023; Tang et al., 2024), bypass the need for an explicit reward model and directly train the LMs on the human preference data. Reinforcement Learning with AI Feedback (Pang et al., 2024) 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. (2025) study whether standard RLHF remains effective under unreliable feedback.

3. Preliminaries

3.1. LLM Alignment with Human Preferences

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

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

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.

The optimal solution to Eq.(1) results in a duality between the reward function $r(x, y)$ and the language model $\pi_\theta(y|x)$:

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

where $Z(x) = \sum_y \pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$ is the normalization factor.

3.2. Offline Fine-Tuning Methods for Reward Maximization

Offline Supervised Methods. Directly optimizing objective 1 require repeated sampling, which can be computationally expensive. This alternative class of methods, including RAFT (Dong et al., 2023) and RWR (Peters and Schaal, 2007), 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:

$$\max_{\pi_\theta} \mathbb{E}_{(x,y_{1,\cdots K})\sim\mathcal{D}_{\text{off-sup}}}\left[\log\pi_\theta(y_i|x)\cdot F\left(x,y_i|y_{1,\cdots,K}\right)\right]$$

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 (Burns et al., 2024), 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}. textbf(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 (Rafailov et al., 2023; Levine, 2018; Haarnoja et al., 2017):

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

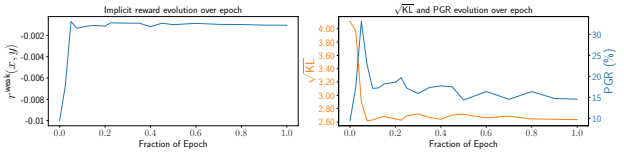

Figure 2: **Learning from weak supervision** as reward maximization. **Left:** the strong model π_θ learns to maximize the implicit reward $r^{\text{weak}}(x, y) = \beta \log \pi^{\text{weak}}_{\text{align}}(y|x)$ - $\beta \log \pi^{\text{weak}}_{\text{ref}}(y|x)$. **Right:** the strong model also learns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

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

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:*

$$\min_{\pi_\theta} KL\left(\pi^{weak}||\pi_\theta\right) \quad (3)$$
$$s.t\ \pi^{weak} = \arg\min_{\pi} KL\left(\pi||\pi^{EBM}\right)$$

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

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:*

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[r^{weak}(x, y)\right] - \lambda KL(\pi_\theta||\pi^{weak}_{ref}) \quad (4)$$

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,

$$\min_{\pi_\theta} \text{KL}\left(\pi_\theta||\pi^{\text{weak}}\right), \quad (5)$$

is equivalent to maximizing the KL-constrained reward objective in Eq. (4). □

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:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[r^{\text{weak}}(x, y)\right] - \lambda\text{KL}(\pi_\theta||\hat{\pi}) \quad (6)$$

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. (2024) 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., $\hat{\pi}(y|x) = \pi^{\text{strong}}_{\text{ref}}(y|x)$. This design choice serves an important goal: to encourage the learned policy π_θ to remain close to the initial strong reference model $\pi^{\text{strong}}_{\text{ref}}$, thereby facilitating the elicitation of the student's prior knowledge while simultaneously incorporating assessment from the weak teacher. Similar to (Burns et al., 2024), 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 $\pi^{\text{strong}}_{\text{ref}}$ to generate task-relevant outputs, as $\pi^{\text{strong}}_{\text{ref}}$ is not trained to follow instructions.

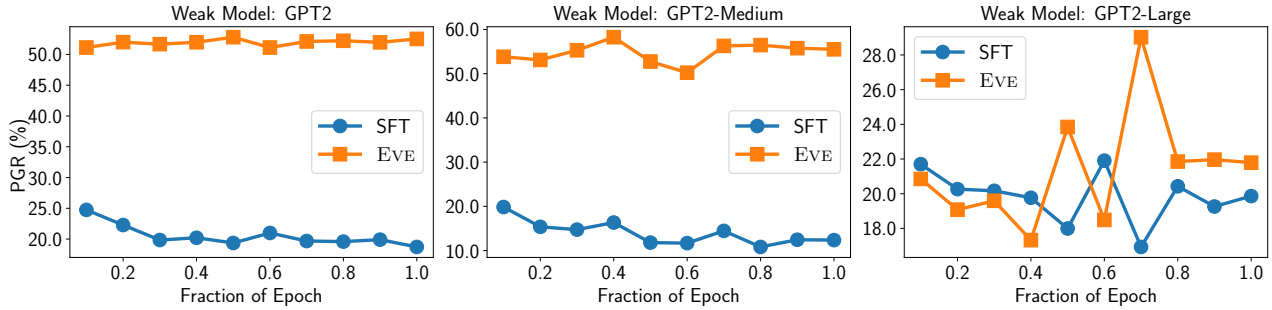

Figure 3: Evolution of PGR (%). We observe clear signs of overfitting to the weak teacher's errors well before finishing a single epoch. Notably, when there is a large gap between the strong student and the weak teacher, the student reaches its best performance within the first 10% of the epoch. EVE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

Optimization. Directly optimizing the objective in Eq. (6) can incur significant computational costs, as it requires repeated sampling from the strong student π_θ inside the training loop (Rafailov et al., 2023). Following prior work (Rafailov et al., 2023; Peters and Schaal, 2007; Peng et al., 2019), 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} (Rafailov et al., 2023). 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* (Owen, 2013). 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:

$$F(x, y^i | y^{1,\cdots,K}) = \frac{K \cdot \exp\left(f(x,y^i)\right)}{\sum_{k=1}^K \exp\left(f(x,y^k)\right)} \tag{7}$$

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 **controlled-summarization** tasks:

Setup. We choose the representative Reddit TL;DR summarization (Stiennon et al., 2020) dataset and follow the synthetic setup from (Gao et al., 2023; Zhou et al., 2024; Rafailov et al., 2023), 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 (Radford et al., 2019) (GPT2-Base/Medium/Large) as weak teachers and a more advanced LLama-3.2-3B model (MetaAI, 2024a;b) as the strong student. The weak model π^{weak} is the aligned model with DPO (Rafailov et al., 2023) from $\mathcal{D}_{\text{golden}}$.

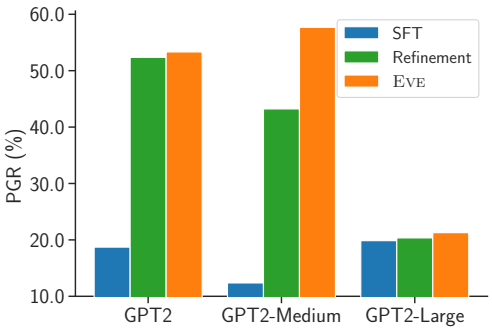

Figure 4: PGR (%) of SFT, Refinement and EVE.

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** (Somerstep et al., 2024; Yang et al., 2024) – 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.

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 (Park et al., 2024).

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

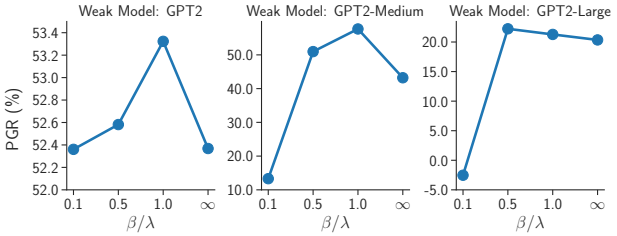

Figure 5: PGR (%) of various β/λ ratios in EVE's objective.

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.

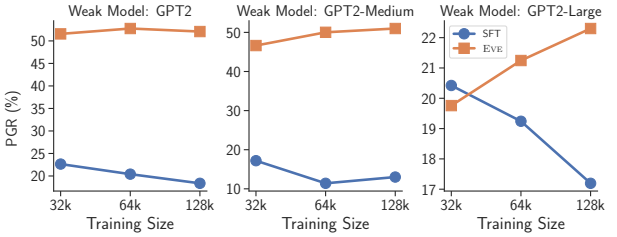

Figure 6: Scaling the training size (32k, 64k and 128k) in EVE and SFT (trained for one epoch). EVE shows notable improvement as the training size increases, while SFT suffers from overfitting.

6. 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.

Impact Statement

Our work demonstrates a positive societal impact with better alignment with human values, including helpfulness and harmlessness. We do not expect any negative societal impacts directly resulting from the contributions presented in our paper.

References

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.

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.

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.

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.

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.

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.

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.

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.

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/.

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

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

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

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

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

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.

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

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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)*, 2024a.

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*, 2024b.

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/.

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.

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.

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

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.