

821 A Societal Impacts

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

825 A.1 Computing Resources Specification

826 We train and evaluate our models using 4xNVIDIA H100 GPUs. All evaluations are judged by
827 "gpt-4o-mini", with random positional flips to avoid position bias.

828 B Further Details on the Experimental Setup

829 B.1 Gold-reward training details

830 We follow the synthetic setup in which we use the gold reward models to play the roles of humans
831 and provide preference feedback [10, 47].

832 The golden reward model is initialized from Llama-3.1-8B. We first pool together the human prefer-
833 ence Reddit TL;DR summarization dataset. We, then, apply SFT on the SFT split in Reddit TL;DR
834 dataset. We then fine-tune the golden reward on the preference dataset. For both SFT and reward
835 modeling phases, we use a batch size of 128, a learning rate of $1e-6$ over one epoch with a cosine
836 learning rate schedule with 150 warm-up steps. The golden reward model, then, generates synthetic
837 preferences on the original dataset with $p(y_1 \succ y_2 \mid x) = \sigma(r_{\text{gold}}(x, y_1) - r_{\text{gold}}(x, y_2))$, where σ is
838 the sigmoid function. The gold reward models achieve high test accuracies on summarization tasks
839 with 75.8% accuracy, showing a strong correlation with human preferences.

840 B.2 Direct Alignment Algorithms (DAAs) for weak models training setups

841 We apply DAAs on the synthetic golden preferences $\mathcal{D}_{\text{golden}} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$, which involves
842 two stages: Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) [30]. During
843 SFT, the weak models are trained on *chosen responses only*. During DPO, we use standard setting of
844 $\beta = 0.1$. For both phases, we use a batch size of 128, a learning rate of $1e-6$, and a cosine learning
845 rate scheduler with a warmup of 150 steps.

846 B.3 Details of estimating KL divergence

847 We can construct an unbiased estimate of $\text{KL}(\pi_\theta \parallel \pi_{\text{ref}})$ by sampling. Specifically, we first sample
848 N prompts $\{x^{(i)}\}_{i=1}^N$ from the evaluation set, and then for each prompt $x^{(i)}$, we sample a response
849 $y^{(i)} \sim \pi_\theta(\cdot \mid x^{(i)})$ from the learned policy π_θ . We can estimate the KL divergence at the sequence
850 level as follows:

$$\frac{1}{N} \sum_{i=1}^N \log \pi_\theta(y^{(i)} \mid x^{(i)}) - \log \pi_{\text{ref}}(y^{(i)} \mid x^{(i)})$$

851 However, this estimation has high variance and can be negative. Therefore, we use the following
852 unbiased estimator [32]:

$$\frac{1}{N} \sum_{i=1}^N \left(\frac{\pi_{\text{ref}}(y^{(i)} \mid x^{(i)})}{\pi_\theta(y^{(i)} \mid x^{(i)})} - 1 \right) + \log \frac{\pi_\theta(y^{(i)} \mid x^{(i)})}{\pi_{\text{ref}}(y^{(i)} \mid x^{(i)})}$$

853 B.4 Simulating Weak Supervision on Reward Modeling

854 **Models:** The reward model is first initialized from the supervised fine-tuned stage in RLHF, we
855 remove the unembedding layer of the model and add a linear head to output a scalar value, which
856 represents the reward for model completion.

857 **Hyperparameters:** We use a learning rate of $1e-6$, and a cosine learning rate scheduler for training
858 the reward model with a batch size of 64 and a warmup of over 150 steps.

859 **Weak labels:** We train a proxy reward model on half of golden preference dataset $\mathcal{D}_{\text{golden}}$ that is
860 labeled by the golden reward model and then generate weak labels on the other half to create a weak
861 preference data $\mathcal{D}_{\text{pref}}$. The weak label preference dataset is used to train a new reward model $\tilde{r}(x, y)$
862 for RL fine-tuning stage.

863 B.5 Detailed Calculation of Performance Gap Recovered (PGR)

864 To evaluate the performance \mathcal{P} of the LM policy π , we use the golden reward model to score π
865 generated responses. Specifically, given a policy, its performance is defined as follows:

$$\mathcal{P}(\pi) = \mathbb{E}_{x \sim \mathcal{D}_{\text{test}}, y \sim \pi(\cdot|x)} [r_{\text{gold}}(x, y)] \approx \frac{1}{N} \sum_{i=1}^N r_{\text{gold}}(x, y)$$

866 where $\mathcal{D}_{\text{test}}$ is an evaluation set of prompts. Performance Gap Recovered (PGR) of a W2S policy π_{θ}
867 is calculated by:

$$\text{PGR} = \frac{\mathcal{P}(\pi_{\theta}) - \mathcal{P}(\pi^{\text{weak}})}{\mathcal{P}(\pi^{\text{strong}}) - \mathcal{P}(\pi^{\text{weak}})}$$

868 For all experiments, we set $N = 256$, similar to prior works [30, 31].

869 B.6 Prompt Template for Refinement Sampling from Strong Models

870 **Controlled-summarization:** We use 2-shot prompting to elicit strong student knowledge. To be
871 more specific, we randomly sample 2 prompts on a held-out set of prompts dedicated only to training
872 the strong student. These 2 prompts are input to the weak aligned model to generate 2 responses. We
873 then use the two responses as a demonstration using the following format:

{exemplar[1].prompt}TL;DR: {exemplar[1].response}
 {exemplar[2].prompt}TL;DR: {exemplar[2].response}
 {prompt}TL;DR:

874 **Instruction following.** For the instruction following task, we ask the strong student to refine the
875 weak teacher’s response to align with the helpfulness objective with the following format:

User: {exemplar[1].prompt}
 Weak Assistant: {exemplar[1].response}
 Can you make the response more helpful and coherent?
 Strong Assistant:

876 C A Bayesian Interpretation of W2S Generalization

877 Burns et al. [4] saliency interpretation can essentially be viewed as a Bayesian inference problem.
878 Intuitively, Bayesian inference is the problem of updating a distribution to conform to new evidence.

879 In W2S generalization setup, we update the strong student π_{θ} , which is initialized from a prior $\pi_{\text{ref}}^{\text{strong}}$,
880 that represents its knowledge about the tasks to conform with the evidence provided by the weak
881 teacher data. The weak teacher data provides evidence about the intended tasks that we want the
882 strong model to solve but also contains systematic errors of the weak teacher. Then, we can define
883 the posterior distribution:

$$\begin{aligned} \pi^*(y|x) &= \frac{1}{Z(x)} \pi^{\text{weak}}(y|x) \pi_{\text{ref}}^{\text{strong}}(y|x) \\ &= \frac{1}{Z(x)} \exp(r'(x, y)/\beta) \pi_{\text{ref}}^{\text{strong}}(y|x) \end{aligned}$$

884 Interestingly, this coincides with the optimal solution of the KL-constrained reward maximization
885 objective where the reward is defined by $r'(x, y) = \log \pi^{\text{weak}}(y|x)$. Intuitively, rather than directly
886 training the strong student to imitate the weak teacher can also lead to imitating the errors of the
887 teacher. The posterior formulation avoids this failure mode by reinforcing the strong student to
888 remain close to its initial knowledge while still learning the tasks effectively.

D Mathematical Derivations

Lemma D.1. Consider an optimal policy such that there exists a minimizer to the following optimization problem:

$$\begin{aligned} \min_{\pi_\theta} KL(\pi^{weak} || \pi_\theta) \\ s.t. \pi^{weak}(y|x) = \arg \min_{\pi} KL(\pi || \pi^{EBM}) \end{aligned}$$

where $\pi^{EBM}(y|x) \propto \pi_{ref}^{weak}(y|x) \exp(r^{weak}(x, y)/\beta)$. Then, the optimal policy π^* for the above objective is when $\pi^*(y|x) = \pi^{weak}(y|x) = \pi^{EBM}(y|x)$.

Proof. The optimal solution to the KL divergence in the lower-level objective is minimized when $\pi^{weak} = \pi^{EBM}$. substitute this optimal policy in the lower-level objective to the upper-level objective:

$$\min_{\pi_\theta} KL(\pi^{EBM} || \pi_\theta)$$

The solution to this optimization problem is achieved when $\pi_\theta = \pi^{EBM}$. Which concludes the proof. \square

Theorem D.2. The optimal solution to W2S generalization is equivalent to the optimal solution in the following objective:

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}_{weak}, y \sim \pi_\theta(\cdot|x)} [r^{weak}(x, y)] - \beta KL(\pi_\theta || \pi_{ref}^{weak}) \quad (9)$$

Proof: Following Appendix A.1 of [30], we have:

$$\begin{aligned} & \max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta} [r(x, y)] - \beta KL(\pi_\theta || \pi_{ref}^{weak}) \\ &= \max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta} \left[r(x, y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}^{weak}(y|x)} \right] \\ &= \min_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)} \left[\log \frac{\pi_\theta(y|x)}{\pi_{ref}^{weak}(y|x)} - \frac{1}{\beta} r(x, y) \right] \\ &= \min_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)} \left[\log \frac{\pi_\theta(y|x)}{\frac{1}{Z(x)} \pi_{ref}^{weak}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)} - \log Z(x) \right] \end{aligned} \quad (10)$$

where the normalization factor is:

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

Given an EBM policy defined in Lemma 4.2, we can re-organize objective [10] we have:

$$\min_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)} \left[\log \frac{\pi_\theta(y|x)}{\pi^{EBM}(y|x)} - \log Z(x) \right] \quad (11)$$

$$= \min_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}} [KL(\pi_\theta(\cdot|x) || \pi^{EBM}(\cdot|x)) - \log Z(x)] \quad (12)$$

Note that $Z(x)$ does not depend on π_θ . Therefore, the RLHF objective is maximized with respect to π_θ if and only if $\pi_\theta(y|x) = \pi^{EBM}(y|x)$. By Lemma D.1, we know that the optimal solution to W2S generalization problem is achieved when $\pi_\theta(y|x) = \pi^{EBM}(y|x)$. Therefore, we have proved that the optimal solution to the RLHF objective and weak-to-strong generalization is equivalent.

D.1 Utilizing Strong Student's Knowledge Only for Reward Evaluation

EVE must sample or elicit responses from the strong student during training, to leverage its prior knowledge and balance learning from weak supervision. On the other hand, we could instead use the

strong student for reward evaluation of weak supervision data $\mathcal{D}_{\text{weak}}$. Specifically, this results in the following reward function:

$$r^{\text{strong}}(x, y) = \log \pi_{\text{ref}}^{\text{strong}}(y|x) - \log \pi_{\text{ref}}^{\text{weak}}(y|x)$$

and the following optimization objective:

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

Intuitively, this objective can be interpreted as a weighted maximum likelihood objective that weighs samples generated by the weak teacher using the implicit reward function of the strong reference model; this allows the training process to scale down undesirable knowledge from the weak reference model π_{ref} .

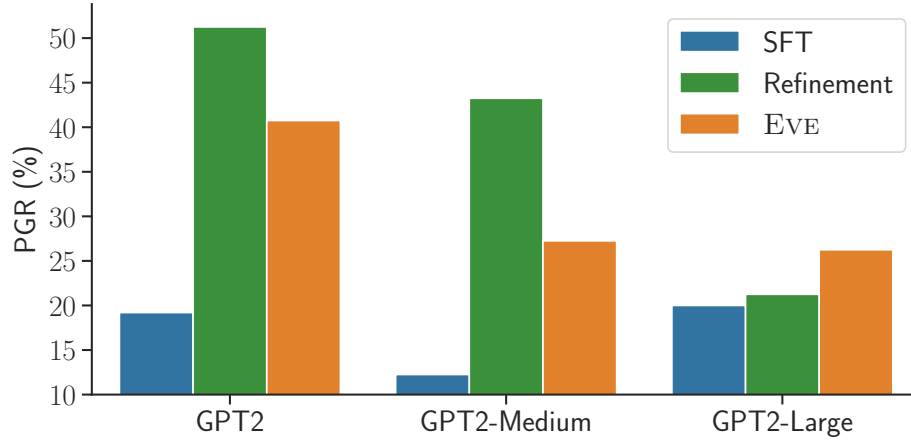


Figure 8: PGR (%) when using the strong student as supervision signal on weak supervision data. We observe that using the strong student as a supervision signal can provide performance improvement compared to SFT.

Fig. 8 illustrates the performance of this approach and the baselines. As can be observed, using the strong student for reward assessment significantly outperforms SFT across different model sizes. This result highlights the substantial benefits of leveraging the strong student in W2S learning. On the other hand, this approach does not outperform Refinement when the weak model is GPT2-Large. Nevertheless, Refinement approaches are more computationally expensive, as they rely on the strong student to “refine” (involves a generation step) data generated by the weak teacher, while this approach utilizes the strong student to only “evaluate” the weak supervision samples.

D.2 Comparison with DPO in Controlled Summarization

Tajwar et al. [36] show that contrastive learning methods, which incorporate negative gradients, outperform offline supervised methods. This is because the negative gradient allows the policy to deviate significantly from its initial distribution when seeking the high reward function, whereas offline supervised methods are unable to guide the policy far from its initial distribution. However, it remains unclear whether this negative gradient provides similar benefits compared to offline supervised approaches in W2S generalization with potentially unreliable supervision.

Fig. 9 shows the Win rates across different KL-budgets, which corresponds to different values of the KL-regularization hyperparameter, of EVE and DPO. As can be observed, across different weak models, DPO suffers significantly from reward over-optimization, where increasing the KL budget results in a decline in model performance [31]. In the context of W2S generalization, this problem is significant because we cannot easily find an early stopping criterion for training the strong student. In contrast, EVE, an offline supervised approach, does not diverge much from its initial distribution, making it significantly more robust to over-optimization phenomenon across different regularization hyperparameters.

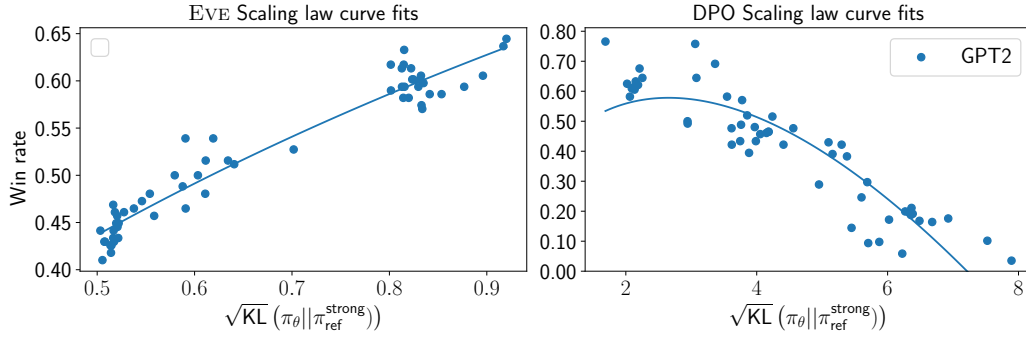


Figure 9: Trade-off between Win-rate and KL-between DPO and EVE methods with varying regularization strength.

939 E Sample Refined Responses Across Weak Teachers

Prompt	SUBREDDIT: r/Advice TITLE: How to help repay brother? POST: So my older brother is going to college and he isn't necessarily nice (he's pretty mean and doesn't talk much). He surprised me today by giving me 3 Xbox One games. However, 2 of these games are almost brand new and one is brand new. They are all popular and expensive games and I want to repay him but I don't play Xbox One anymore. I'm still grateful but should I sell them and profit about \$150 (I'm only 13 so a lot of money) or should I keep it for sentimental value? I know it sounds bad for me wanting to sell but he's been rude, disrespectful and even abusive once so I don't know what to do. Also should I give it back? Is it the right thing to do?
GPT2	TL;DR: Brother gives me 3 Xbox One games and I'm really grateful but should I sell them and profit about \$150 (I'm only 13 so a lot of money) or should I keep it for sentimental value? I don't play Xbox One anymore. I'm still thankful but should I sell them and profit about \$150 (I'm only 13 so a lot of money) or should I keep it for sentimental value? I know it sounds bad for me wanting to sell but he's been rude, disrespectful and even abusive once so I don't know what to do. Also should I give it back? Is it the right thing to do
Strong student	TL;DR: My brother gave me 3 Xbox One games that are almost brand new and expensive. He is rude and has been abusive. Should I sell it and give him \$150 or should I keep it for sentimental value?

Table 1: Sample prompt, responses from GPT2 teacher response and the strong student after refining.