WHEN WEAK LLMS SPEAK WITH CONFIDENCE, PREFERENCE ALIGNMENT GETS STRONGER

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference alignment is an essential step in adapting large language models (LLMs) to human values, but existing approaches typically depend on costly human annotations or large-scale API-based models. We explore whether a weak LLM can instead act as an effective annotator. We surprisingly find that selecting only a subset of a weak LLM's highly confident samples leads to substantially better performance than using full human annotations. Building on this insight, we propose ***C**onfidence-**W**eighted **P**reference **O**ptimization* (CW-PO), a general framework that re-weights training samples by a weak LLM's confidence and can be applied across different preference optimization objectives. Notably, the model aligned by CW-PO with just 20% of human annotations outperforms the model trained with 100% of annotations under standard DPO. These results suggest that weak LLMs, when paired with confidence weighting, can dramatically reduce the cost of preference alignment while even outperforming methods trained on fully human-labeled data.

1 INTRODUCTION

Large language models (LLMs) are typically developed through three stages: large-scale pre-training with next-token prediction, supervised fine-tuning (SFT), and preference alignment. While pre-trained and SFT models can generate coherent and task-oriented text, their outputs often remain misaligned with human expectations, exhibiting issues such as bias, factual errors, or unsafe content. Preference alignment addresses this gap by steering models toward desirable behaviors such as helpfulness, harmlessness, and truthfulness, thereby improving their reliability and trustworthiness in real-world applications.

Preference alignment methods, such as reinforcement learning from human feedback (RLHF) (Christiano et al., 2017) or direct preference optimization (DPO) (Rafailov et al., 2023), rely on a prompt paired with two candidate responses (x, y_1, y_2), where annotators judge which response better fits a given criterion. Since candidate responses y_1 and y_2 can be easily generated through LLM prompting, collecting triplets is straightforward; however, obtaining human preference data is expensive and time-consuming. Moreover, collected datasets are prone to noise due to the subjectivity of human judgements, which vary across contexts and annotators (Bai et al., 2022; Ouyang et al., 2022; Cui et al., 2023; Gao et al., 2024). Thus, obtaining high-quality preference datasets remains a challenge.

An alternative is to use large-scale API-based LLMs as annotators (*e.g.*, ChatGPT) (Dubois et al., 2023; Ye et al., 2023; Kim et al., 2023; Lee et al., 2023), but these still incur substantial computational and financial costs. Interestingly, recent work (Tao & Li, 2025) has shown that even weak LLMs (*e.g.*, OPT-125M (Zhang et al., 2022)), when trained on a small amount of human data, can serve as annotators to align stronger models – sometimes even reaching or surpassing performance achieved with human-labeled supervision. However, they treat weak-model predictions directly as preference annotations, raising the question of *how to more effectively leverage them for alignment*.

In this work, we propose ***C**onfidence-**W**eighted **P**reference **O**ptimization* (CW-PO), a highly effective preference alignment approach that requires minimal human supervision for alignment and is compatible with different preference optimization methods. CW-PO is motivated by a key observation that a subset of high-confidence predictions from a weak LLM are more effective for aligning

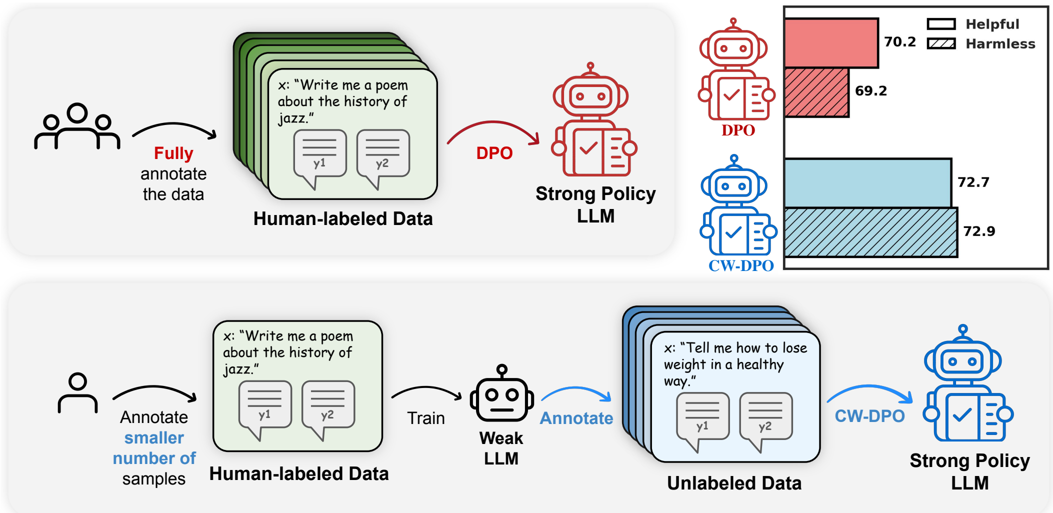

Figure 1: Overall pipeline of our setting. *Top*: Conventional DPO (Rafailov et al., 2023). For each triplet consisting of a prompt x and two candidate responses (y_1, y_2), human annotators provide preference labels, and the policy model is aligned with these labels using DPO. *Bottom*: CW-DPO framework. A weak LLM is first trained as a preference annotator using a subset of human-labeled triplets. It is then applied to annotate the remaining large-scale data, which is subsequently trained with CW-DPO. The bars on top right report Gold Reward Accuracy for standard DPO with human-labeled data (red) and for CW-DPO (blue) on the ANTHROPIC HH-RLHF. CW-DPO uses only 30% compared to DPO, which uses fully human-annotated dataset. OPT-125M and OPT-1.3B are used as the weak and strong models, respectively.

stronger LLMs than using fully human-labeled data. Leveraging this insight, CW-PO reweights samples in the preference optimization objective according to the confidence of a weak LLM. CW-PO offers three main advantages:

- **High performance**: We show that with a small amount of human-annotated data, a weak LLM can be trained into an effective preference annotator. As a concrete instantiation, we apply CW-PO to the Direct Preference Optimization (DPO) loss (Rafailov et al., 2023), yielding CW-DPO. We show that with 30% annotations of the dataset, CW-DPO outperforms the model trained with the full 100% of the human annotations (Figure 1). Notably, CW-DPO remains more effective even with *just 20% annotations*. Moreover, CW-PO substantially outperforms the direct use of weak model annotations for supervision, the approach employed by Tao & Li (2025).
- **Low computational cost**: We use weak annotators with *fewer than 0.5B parameters* and show that even a lightweight 125M model can be highly effective. This makes obtaining annotations far cheaper than relying on humans and far more efficient than using large API-based LLMs such as ChatGPT, with substantial savings in both inference time and memory.
- **Extensibility**: Once trained on a small amount of human-labeled data, a weak LLM annotator can be repeatedly reused with CW-DPO for preference data annotation. This is highly practical because generating triplets (x, y_1, y_2) via prompting an LLM is straightforward, whereas reliably annotating them remains a major challenge.

2 Problem Statement and Preliminaries

2.1 Problem statement

We aim to align a strong LLM under the supervision of a weaker LLM. We follow the setup of Tao & Li (2025), which fine-tunes the weak model on a subset of preference triplets with human annotations and then uses its predictions to label the remaining data. Based on this setup, we define our problem as follows:

Definition 1 (Preference Data). *Let $\mathcal{D}_{preference}$ denote a collection of tuples, each consisting of a single prompt and two corresponding responses, along with an annotation indicating which response is more preferable.*

$$\mathcal{D}_{preference} = \{(x, y^+, y^-) \mid x \in \mathcal{X},\ y^+, y^- \in \mathcal{Y},\ y^+ \succ y^-\}, \tag{1}$$

where $\mathcal{X}$ denotes the space of prompts, $\mathcal{Y}$ denotes the space of candidate responses, and $y^+ \succ y^-$ indicates that y^+ is preferred over y^- for prompt x according to human preference.

We are provided with a smaller labeled subset $\mathcal{D}_{\text{labeled}} \subset \mathcal{D}_{\text{preference}}$ containing human annotations (*e.g.*, 34,000 samples, corresponding to 20% of ANTHROPIC HH-RLHF (Bai et al., 2022) dataset), and a large unlabeled subset $\mathcal{D}_{\text{unlabeled}}$, such that $\mathcal{D}_{\text{labeled}} \cup \mathcal{D}_{\text{unlabeled}} = \mathcal{D}_{\text{preference}}$.

2.2 PRELIMINARIES

Tao & Li (2025) first fine-tune a weak LLM π_w on $\mathcal{D}_{\text{labeled}}$ to predict preference labels. The weak LLM is then applied to $\mathcal{D}_{\text{unlabeled}}$ to produce preference annotations:

$$\hat{\mathcal{D}} = \{(x, y^+, y^-) \mid y^+ \succ_{\pi_w} y^-\}, \tag{2}$$

where $y^+ \succ_{\pi_w} y^-$ indicates that π_w predicts y^+ to be preferable over y^-.

Finally, the weakly-labeled pairs—annotated by the weak LLM—are used to align the strong policy π_s via the preference optimization objective.

Definition 2 (Preference Optimization Objective). *Given a dataset of annotated triplets $\hat{\mathcal{D}} = \{(x, y^+, y^-)\}$, the goal of preference optimization is to align a policy model π_s such that it assigns a higher likelihood to preferred responses. This is formalized as the expected loss:*

$$\mathcal{L}_{PO}(\pi_s; \hat{\mathcal{D}}) = \mathbb{E}_{(x,y^+,y^-)\sim\hat{\mathcal{D}}} \left[\ell(\pi_s; x, y^+, y^-) \right], \tag{3}$$

where $\ell(\cdot)$ denotes a generic preference optimization (PO) loss function, such as DPO (Rafailov et al., 2023), IPO (Azar et al., 2024), rDPO (Chowdhury et al., 2024), or other variants. The details of these loss functions are provided in Appendix B.

The objective is to align π_s more faithfully to human preferences by leveraging data annotated by the computationally inexpensive weak LLM π_w.

In this scenario, Tao & Li (2025) adopt DPO as the preference optimization loss and show that even weak LLMs can serve as effective annotators for aligning stronger models, at times matching or surpassing the performance of human supervision. Building on this finding, we follow the setting of Tao & Li (2025) to explore ***how weak LLMs can be more effectively leveraged to align a strong model.***

Note that this scenario is highly practical, as a large volume of triplets (x, y_1, y_2) can be obtained with minimal effort. For any given prompt, generating two or more diverse responses is straightforward via standard prompting techniques in modern LLMs. Moreover, human-annotated datasets, which can be used as $\mathcal{D}_{\text{labeled}}$ for alignment criteria such as helpfulness and harmfulness, are already available, including ANTHROPIC HH-RLHF.

3 CONFIDENCE-WEIGHTED PREFERENCE OPTIMIZATION

3.1 EXPLORATION ON WEAK LLM CONFIDENCE

We find that leveraging the confidence predicted by a weak LLM can substantially improve the alignment of a stronger model. Using the pairwise ANTHROPIC HH-RLHF dataset (Bai et al., 2022), we compute, for each triplet (x, y_1, y_2) in $\mathcal{D}_{\text{unlabeled}}$, the absolute difference between the weak model's predictions for the two candidate responses, *i.e.*, $|\pi_w(x, y_1) - \pi_w(x, y_2)|$, which intuitively reflects the weak model's confidence in distinguishing the preferred response[1]. We then apply thresholding to select the top-N% of samples with the highest confidence scores. For example, with the top 30%,

[1] A detailed explanation of the weak model's training procedure is provided earlier in Section 3.2.

we use the subset consisting of the 30% most confident samples from $\hat{\mathcal{D}}$'. For the experimental results presented in Figure 2, we use two subsets of the HH-RLHF dataset, "Harmless" and "Helpful", and their concatenation is denoted as "HH-RLHF". Additionally, the *Human* bars correspond to the results of LLMs trained on the human-annotated dataset. Notably, even with fewer training samples, decreasing the confidence threshold consistently improves performance. For "Helpful", the trend is less gradual but still striking: training on only the top 30% most confident samples achieves the highest reward accuracy by a clear margin. These results extend the finding of Tao & Li (2025)—that weak-LLM annotations can mostly surpass human annotations (100% is better than *Human* in Figure 2 for "Harmless" and "HH-RLHF" datasets)—by showing that combining weak LLMs with their prediction confidence enables *even more effective* preference alignment than naive usage of weak-LLM annotations. This naturally raises the next question: ***How can we systematically incorporate this crucial observation into the alignment paradigm?***

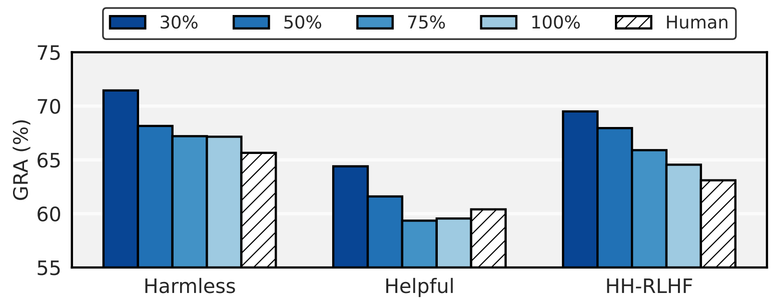

Figure 2: Alignment with the top-N% most confident samples. Gold reward accuracy (GRA) is reported for the trained strong models. We consider (OPT-125M $\rightarrow$ OPT-1.3B) and (Qwen-0.5B $\rightarrow$ Qwen-7B) as weak–strong model pairs. The graph shows the average GRA for two models. Here, 100% denotes using the weak LLM directly for annotation. Further details of the results are provided in Appendix C.1.

3.2 Confidence-Weighted Preference Optimization

We introduce *Confidence-Weighted Preference Optimization* (CW-PO), a new alignment framework that incorporates weak-LLM confidence scores into the standard PO objective (Equation 3). Intuitively, as motivated in Section 3.1, it is preferable to assign greater weight to samples with higher confidence and smaller weight to those with lower confidence. To achieve this, we propose a three-step framework: (*i*) We train a weak LLM as a preference annotator; (*ii*) The trained weak LLM is used to generate preference labels for unlabeled prompt-response pairs, selecting the preferred and rejected responses based on their predicted scores; and (*iii*) We align a stronger LLM by introducing a confidence-based weight into the PO objective, which prioritizes high-confidence samples for more effective alignment. We next describe each of the steps in detail.

***(i)* Constructing a preference annotator.** For the weak model, we used its pretrained backbone, bypassed the last layer, and added a scalar output layer. We then optimized the entire model. Using the pretrained backbone allows us to transfer the knowledge from the weak LLM to a preference annotation task, requiring only a small amount of data to achieve an accurate annotator.

The Bradley-Terry (BT) (Bradley & Terry, 1952) model provides a principled way to connect reward modeling with preference learning. It models the probability of one option being preferred over another as:

$$p(y^+ \succ y^- \mid x) = \sigma(\pi_w(x, y^+) - \pi_w(x, y^-)), \tag{4}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function, and $\pi_w : (\mathcal{X}, \mathcal{Y}) \rightarrow \mathbb{R}$ is the weak LLM's scoring function (logit) for a given response to a prompt. The model is then optimized by minimizing the negative log-likelihood of the human preference data[2]:

[2]Whereas we let the final layer of the weak LLM perform preference classification, Tao & Li (2025) do not modify the LLM's output; instead, they compute an implicit reward based on response generation and use it as a pseudo label. We discuss this in detail in Appendix G.

$$\mathcal{L}_{\text{weak}} = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}_{\text{labeled}}}\Big[\log\sigma(\pi_w(x,y^+) - \pi_w(x,y^-))\Big]. \tag{5}$$

This objective encourages the weak LLM to relatively assign higher scores to preferred responses and lower scores to dispreferred ones.

***(ii)* Generating preference labels.** After fine-tuning, the weak LLM π_w is applied to unlabeled pairs to determine preference labels. Given a prompt x and two (unlabeled) candidate responses (y_1, y_2), we define the chosen and rejected responses according to the weak model's scoring function:

$$y^+ = \arg\max_{y\in\{y_1,y_2\}} \pi_w(x,y), \quad y^- = \arg\min_{y\in\{y_1,y_2\}} \pi_w(x,y). \tag{6}$$

That is, the response with the higher weak-model score is treated as the *chosen* response y^+, while the other is treated as the *rejected* response y^-. According to Equation 2, this procedure produces the weakly-labeled preference dataset $\hat{\mathcal{D}}$.

***(iii)* Aligning a strong large language model.** Building on PO (Equation 3), we propose CW-PO, which introduces a confidence-based weight into the loss:

$$\mathcal{L}_{\text{CW-PO}} = \mathbb{E}_{(x,y^+,y^-)\sim\hat{\mathcal{D}}}\left[\mathcal{C}(x,y^+,y^-)\cdot\ell(\pi_s;x,y^+,y^-)\right], \tag{7}$$

where $\mathcal{C}(x,y^+,y^-)$ is the confidence score, defined as the prediction margin between the weak model's scores for the preferred and rejected responses:

$$\mathcal{C}(x,y^+,y^-) = 2\cdot(\sigma(\pi_w(x,y^+) - \pi_w(x,y^-)) - 0.5), \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function. Since $\pi_w(x,y^+) - \pi_w(x,y^-) \geq 0$ (y^+ has higher predicted score than y^-), σ outputs $[0.5, 1]$. Subtracting 0.5 and scaling by 2 normalize it to $[0, 1]$. A detailed analysis of this choice of weighting is presented in Appendix H.

CW-PO can be instantiated with different PO strategies. By applying our framework to DPO, we obtain CW-DPO, which is expressed as:

$$\mathcal{L}_{\text{CW-DPO}} = -\mathbb{E}_{(x,y^+,y^-)\sim\hat{\mathcal{D}}}\left[\mathcal{C}(x,y^+,y^-)\cdot\log\sigma\left(\beta\log\frac{\pi_s(y^+\mid x)}{\pi_{\text{ref}}(y^+\mid x)} - \beta\log\frac{\pi_s(y^-\mid x)}{\pi_{\text{ref}}(y^-\mid x)}\right)\right], \tag{9}$$

where the coefficient $\beta > 0$ controls the degree of allowed divergence from π_{ref}. Note that our framework can also be applied to other POs; additional examples with IPO and rDPO are provided in the Appendix B.1. By scaling the contribution of each training pair with this confidence-based weight, *CW-DPO* prioritizes high-confidence samples, enabling more effective preference alignment. The complete procedure is provided in Algorithm 1 in Appendix D. We will release the code upon acceptance.

4 Experiments

In this section, we empirically validate the effectiveness of CW-PO, supporting our claim of its ability to enhance performance across different preference alignment strategies and model families.

4.1 Experimental Setup

We evaluate the effectiveness of our proposed *CW-PO* framework when it is applied to different preference optimization (PO) methods including three widely used methods: DPO (Rafailov et al., 2023), IPO (Azar et al., 2024), and rDPO (Chowdhury et al., 2024). We compare our framework against human annotation and the method by Tao & Li (2025) under the following settings:

- *Human*: Align π_s on $\mathcal{D}_{\text{unlabeled}}$ using human-provided annotations.
- *Weak LLM-Supervised DPO* (*WS-DPO*) (Tao & Li, 2025): Train the weak model π_w on $\mathcal{D}_{\text{labeled}}$, then align the strong model π_s on $\mathcal{D}_{\text{unlabeled}}$ using π_w's annotations with DPO[3].

[3] Details of this method are provided in Appendix E.

- *CW-DPO*: Train the weak model π_w on $\mathcal{D}_{\text{labeled}}$, then align the strong model π_s on $\mathcal{D}_{\text{unlabeled}}$ using π_w's annotations with CW-DPO.

Note that the alignment data for the strong model is fixed to $\mathcal{D}_{\text{unlabeled}}$, allowing us to directly compare the quality of preference annotations from humans and the weak LLM, as well as to assess how *CW-PO* can further enhance the weak LLM's annotations. Additionally, to ensure a fair comparison, $\mathcal{D}_{\text{labeled}}$ is the same for both Tao & Li (2025) and the *CW-PO* settings unless stated otherwise. Due of the scale of the experiments and the associated computational cost, we report results from a single run, consistent with Tao & Li (2025).

Datasets. We evaluate *CW-PO* with three datasets, ANTHROPIC HH-RLHF (Bai et al., 2022), ULTRAFEEDBACK BINARIZED (UFB) (Cui et al., 2024), and TL;DR (Stiennon et al., 2022). For ANTHROPIC HH-RLHF, we use the "Harmless" and "Helpful" subsets both individually and jointly (denoted as "HH-RLHF"). We preprocess the data by filtering out samples with fewer than 1024 tokens for the TL;DR dataset, and fewer than 512 tokens for the others. In all experiments, the training data is randomly split into 30% for $\mathcal{D}_{\text{labeled}}$ and 70% for $\mathcal{D}_{\text{unlabeled}}$ unless specified otherwise. Further details of the datasets are provided in Appendix F.

Models. We conduct experiments with the OPT (Zhang et al., 2022) and Qwen (Yang et al., 2025) model families. Specifically, we use Qwen2.5-0.5B and OPT-125M, both small-scale models, as weak annotators to provide preference labels. For the strong models, we consider different sizes, all initialized through Supervised Fine-Tuning (SFT) on prompt–chosen response pairs in $\mathcal{D}_{\text{unlabeled}}$. In our approach and in (Tao & Li, 2025), the chosen responses are based on the weak LLM annotations, while for scenarios where π_s is trained on human annotations, the chosen responses are based on the human-provided labels. All models are trained for 5 epochs.

Evaluation metric. We use Gold Reward Accuracy (GRA) as the evaluation metric, which measures how often the score assigned to the aligned model's response by a pretrained reward model is higher than the corresponding score for the SFT model. We use the reward model from (Liu et al., 2025)as the evaluator for HH-RLHF and UFB, and the reward model from (OpenAssistant, 2023) as the evaluator for TL;DR.

4.2 EXPERIMENTAL RESULTS

CW-PO improves alignment performance across different PO methods and model families, compared to *WS-DPO* (Tao & Li, 2025) and the *Human* baseline (Table 1). In particular, *CW-PO* achieves a 5.2% GRA improvement over *WS-DPO* and a 5% improvement over *Human* on average across all experiments. These results underscore two key insights: *(i) CW-PO* makes conventional preference alignment both more effective and cost-efficient. It reduces reliance on expensive human annotations and is more cost-efficient than *WS-DPO* (Tao & Li, 2025) in weak model training (Ta-

Table 1: Results across different preference alignment methods. The reported values are GRA (%). Weak models in WS-DPO and CW-DPO are trained with 30% of human annotated data. Alignment data for the strong model is fixed across all experiments. *CW-PO* columns are highlighted in blue.

OPT-125M → OPT-13B									
	DPO			IPO			rDPO		
Dataset	Human	WS-DPO	CW-DPO	Human	WS-DPO	CW-IPO	Human	WS-DPO	CW-rDPO
HH-RLHF	56.9	56.7	**61.3**	58.2	62.8	**63.5**	55.9	57.6	**63.0**
TL;DR	**57.0**	53.5	56.6	53.3	49.7	**54.6**	54.2	47.7	**61.4**
UFB	61.3	**63.4**	63.1	63.4	61.3	**66.4**	58.9	61.2	**63.7**
Avg.	58.4	57.9	**60.3**	58.3	57.9	**61.5**	56.3	55.5	**62.7**
Qwen2.5-0.5B → Qwen2.5-14B									
	DPO			IPO			rDPO		
Dataset	Human	WS-DPO	CW-DPO	Human	WS-DPO	CW-IPO	Human	WS-DPO	CW-rDPO
HH-RLHF	78.8	**81.4**	80.6	83.4	81.0	**86.8**	81.2	82.2	**86.2**
TL;DR	64.2	64.8	**66.0**	61.8	62.8	**64.2**	67.0	66.4	**68.8**
UFB	78.1	78.3	**80.1**	78.5	77.2	**80.7**	72.4	75.1	**76.8**
Avg.	73.7	74.8	**75.6**	74.6	73.7	**77.2**	73.5	74.6	**77.3**

ble 5); and *(ii) CW-PO* serves as a plug-and-play enhancement for existing PO methods, improving their effectiveness without altering the underlying algorithm.

4.3 Analysis

For further analysis, we conduct additional experiments, using HH-RLHF and CW-DPO unless stated otherwise.

Different student models. We examine whether a weak model can effectively align a range of stronger policy models within the *CW-PO* framework. We vary the strong models across experiments and find that smaller and mid-sized models benefit more by *CW-PO*, whereas gains diminish as the strong model size increases (Table 2).

Table 2: Performance across different student models measured as GRA (%). We use OPT-125M and Qwen2.5-0.5B as the weak models for the OPT and Qwen families, respectively. GRA measures improvement over a model's SFT baseline; thus larger models may not score higher GRA, since stronger baselines leave less room to improve even if absolute performance is higher.

Dataset	Strong	OPT			Strong	Qwen		
		Human	WS-DPO	CW-DPO		Human	WS-DPO	CW-DPO
HH-RLHF	1.3B	**71.5**	66.7	69.9	1.5B	53.4	55.8	**63.3**
	2.7B	55.1	58.5	**60.3**	3B	66.0	63.3	**73.3**
	6.7B	56.1	62.8	**67.6**	7B	71.1	72.0	**75.2**
	13B	56.9	56.7	**61.3**	14B	78.8	**81.4**	80.6
	Avg.	59.9	61.2	**64.8**	**Avg.**	67.3	68.1	**73.1**
TL;DR	1.3B	53.7	44.7	**59.5**	1.5B	51.8	53.7	**60.3**
	2.7B	52.6	51.6	**59.1**	3B	55.0	56.1	**62.7**
	6.7B	57.5	50.2	**57.7**	7B	61.2	60.1	**64.4**
	13B	**57.0**	53.5	56.6	14B	64.2	64.8	**66.0**
	Avg.	55.2	50.0	**58.2**	**Avg.**	58.1	58.7	**63.4**

Comparison to using full human annotations. Unlike the settings in Table 1 and Table 2, where only $\mathcal{D}_{\text{unlabeled}}$ is used to align the strong model, we next investigate whether *CW-DPO* trained exclusively on $\mathcal{D}_{\text{unlabeled}}$ remains competitive when compared against models trained on the full preference dataset (*i.e.*, $\mathcal{D}_{\text{labeled}} \cup \mathcal{D}_{\text{unlabeled}}$) with human annotations. Remarkably, with just 30% of human annotations, *CW-DPO* still outperforms the model trained with 100% of human annotations (Table 3).

Table 3: Comparison between DPO using the fully human-annotated dataset ($\mathcal{D}_{\text{labeled}} \cup D_{\text{unlabeled}}$) and *CW-DPO*. Numbers in parentheses indicate the relative improvement of *CW-DPO* over the human baseline (positive: improvement, negative: drop). *CW-DPO* columns are highlighted in blue.

Dataset	OPT-125M → OPT-1.3B		Qwen2.5-0.5B → Qwen2.5-7B	
	Human (100%)	CW-DPO	Human (100%)	CW-DPO
Harmless	69.2	72.9 (+3.7)	65.7	72.0 (+6.3)
Helpful	70.2	72.7 (+2.5)	58.5	70.8 (+12.3)
HH-RLHF	71.9	69.9 (−2.0)	72.7	75.2 (+2.5)
TL;DR	54.2	59.5 (+5.3)	63.4	64.4 (+1.0)
Avg.	66.4	68.8 (+2.4)	65.1	70.6 (+5.5)

Different split ratios of $\mathcal{D}_{\text{labeled}}$ and $\mathcal{D}_{\text{unlabeled}}$. To evaluate the impact of labeled data size on *CW-PO*, we vary the proportion of $\mathcal{D}_{\text{labeled}}$ while keeping $\mathcal{D}_{\text{unlabeled}}$ fixed. Overall, *CW-PO* tends to outperform *WS-DPO* (Figure 3, Left).

When there is a fixed pool of preference triplets and only a subset is annotated, one can either align directly the policy model on the labeled subset or adopt *CW-PO*. Namely, we can either train π_s directly on $\mathcal{D}_{\text{labeled}}$ using DPO, or first train π_w on $\mathcal{D}_{\text{labeled}}$ and then use its annotations to further train π_s on $\mathcal{D}_{\text{unlabeled}}$. To test robustness in this setting, we compare *CW-DPO* against the baseline of applying DPO directly on $\mathcal{D}_{\text{labeled}}$ under different labeled–unlabeled splits. *CW-DPO* consistently outperforms direct DPO across all split ratios, demonstrating its effectiveness under limited supervision (Figure 3 *Right*). Note that *CW-DPO* with only 20% of the annotations (reported in Figure 3 *Right*) surpasses DPO trained on the fully human-annotated dataset (70.3% vs. 69.7%).

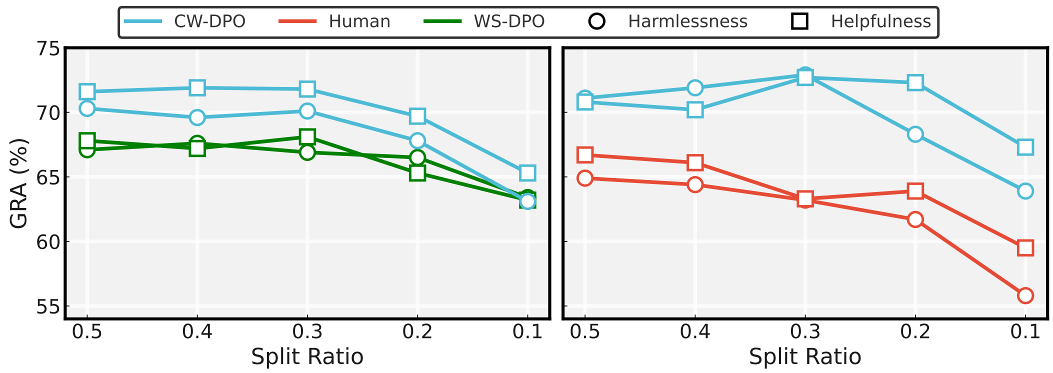

Figure 3: *Left*: GRA when adjusting the proportion of $\mathcal{D}_{\text{labeled}}$ used to fine-tune the weak LLM, while retaining 50% of the data as training for the strong LLM. *Right*: GRA across varying proportions of $\mathcal{D}_{\text{labeled}}$. As the split ratio decreases, the size of $\mathcal{D}_{\text{labeled}}$ decreases and $\mathcal{D}_{\text{unlabeled}}$ increases because the total dataset ($\mathcal{D}_{\text{labeled}} \cup \mathcal{D}_{\text{unlabeled}}$) is fixed.

Comparison to confidence-based filtering. Our *CW-PO* framework is motivated by the observation that filtering the preference alignment data to the most confident examples from a weak model is more effective than leveraging human annotated data. However, filtering based on the confidence is impractical in real-world scenarios because it is difficult to know in advance how to set up the confidence threshold. Nevertheless, we compare *CW-PO* against confidence-based filtering, where only the top-N% most confident samples are retained. We find that *CW-DPO* consistently surpasses the best thresholded setting (30% for HARMLESS/HELPFUL and 40% for HH-RLHF), demonstrating that confidence-based weighting leads to more robust and higher-quality alignment (Table 4). Moreover, we observe two main limitations of confidence-based filtering (Figure 4): (I) the optimal threshold varies across datasets, making it costly and impractical to determine a universal cutoff; (II) setting the threshold too high or too low can dramatically reduce the amount of training data, causing significant performance degradation.

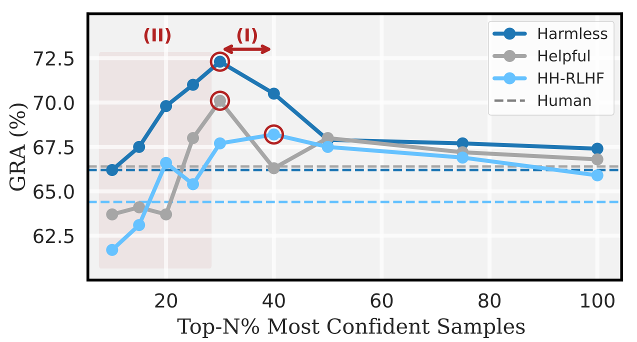

Figure 4: Alignment results across top-N% confidence thresholds.

Table 4: Comparison of confidence-based weighting, *i.e.*, *CW-DPO*, and confidence-based filtering using the top 30% and 40% of samples.

Dataset	Top 30%	Top 40%	CW-DPO
OPT-125M → OPT-1.3B			
HARMLESS	72.3	70.5	**72.9**
HELPFUL	70.1	66.3	**72.7**
HH-RLHF	67.7	68.2	**69.9**
Qwen2.5-0.5B → Qwen2.5-7B			
HARMLESS	70.6	69.1	**72.0**
HELPFUL	58.7	60.2	**70.8**
HH-RLHF	71.3	70.4	**75.2**
Avg.	68.5	67.5	**72.3**

Comparison on the training objective for the weak LLM. We compare the performance of weak LLMs under different training objectives. Using $\mathcal{D}_{\text{labeled}}$ as training data, we benchmark our BT approach against (1) DPO and (2) a two-stage method that first applies supervised fine-tuning (SFT) followed by DPO, as adopted in WS-DPO (Tao & Li, 2025). For evaluation, we use $\mathcal{D}_{\text{unlabeled}}$ with human annotations as a proxy for weak model performance. Across all datasets and both model families, BT consistently achieves the highest reward accuracy while requiring substantially less training time (4,978 vs. 2,450 seconds for 5 epochs) (Table 5). These results highlight that BT

Table 5: Weak models' accuracy.

Model	Dataset	DPO	SFT+DPO	BT
OPT-125M	HARMLESS	55.2	56.3	**69.1**
	HELPFUL	54.1	55.4	**64.2**
	HH-RLHF	50.8	52.1	**63.8**
Qwen-0.5B	HARMLESS	56.1	57.1	**65.3**
	HELPFUL	55.2	56.0	**63.1**
	HH-RLHF	51.4	52.6	**63.2**
Avg.	–	53.8	54.9	**64.8**
Time cost (s)	–	3,319	4,978	**2,450**

not only provides better accuracy but also reduces training cost, making it the most effective and practical choice for training the weak model (See Appendix G for more details).

5 Related Work

Direct preference optimization. Unlike RLHF, DPO directly aligns the policy model with predefined preference data without requiring a reward model (Rafailov et al., 2023). Building on this framework, Identity Preference Optimization (IPO) introduces a regularization term to prevent overfitting (Azar et al., 2024), while Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024) reformulates preference pairs using odds ratios to simplify optimization and improve stability. Simple DPO removes both the reference model and KL penalty, enabling faster and more straightforward training (Meng et al., 2024), and Park et al. (2024) address length bias. Robust DPO (rDPO) (Chowdhury et al., 2024) introduces robustness to noisy preference flips with theoretical guarantees. Contrastive Preference Optimization (CPO) (Xu et al., 2024) frames alignment as a contrastive learning task to maximize the margin between preferred and dispreferred responses. Finally, β-DPO (Wu et al., 2024) proposes dynamically calibrating β at the batch level according to data quality.

Weak-to-strong generalization. Weak-to-strong generalization is a learning paradigm aimed at building superhuman models by leveraging weaker models as proxies for human supervision. The key challenge is that superhuman-level data is often beyond human understanding, making it impossible to provide accurate annotations. Consequently, the focus shifts to how we can effectively elicit the capabilities of a well-pretrained model even under weak supervision (Burns et al., 2024). While our framework also adopts weak-model supervision to align a stronger model, it fundamentally differs from this scenario: in our setting, supervision from a weaker LLM can, in fact, be stronger and even more effective than human annotation.

Large language model-as-a-Judge. Recently, using powerful proprietary LLMs as evaluators for long-form responses has become the de facto standard. Prior work has explored replacing human feedback from AI feedback (Bai et al., 2022), with reinforcement learning from AI feedback (RLAIF) often outperforming human feedback (Lee et al., 2023). Strong LLMs have been used for automatic method evaluation (Dubois et al., 2023) and as examiners that generate questions and assess answers without references (Bai et al., 2023), sometimes decomposing tasks into multiple aspects and criteria for richer evaluation (Saha et al., 2023). Open-source evaluators matching GPT-4's performance with supporting references have also been proposed (Kim et al., 2023). Other efforts include using strong LLMs for automatic low-quality data filtering (Chen et al., 2023) and introducing fine-grained evaluation protocols that break down coarse scores into skill-level assessments (Ye et al., 2023). While these works have relied on strong models' capability (*e.g.*, GPT-4), Tao & Li (2025) demonstrates that even **weaker LLMs** (*e.g.*, OPT-125M) can achieve annotation quality comparable to, or surpassing, that of humans, offering both effectiveness and efficiency.

Building on these insights, this paper further investigates strategies for making more effective use of annotations produced by weak LLMs.

6 Concluding Remarks

In conclusion, we introduced *CW-PO*, a principled framework for leveraging weak LLMs as efficient and scalable preference annotators. By reweighting samples based on annotator confidence, *CW-PO* effectively amplifies the utility of weak-model supervision, achieving strong alignment performance with only a fraction of human-labeled data. Our results demonstrate that even lightweight annotators with fewer than 0.5B parameters can reliably guide much stronger LLMs, offering both substantial computational savings and practical reusability.

Limitation. While CW-PO achieves significant improvements, there may exist other more effective strategies to exploit confidence information for preference alignment. Our main contribution lies in presenting a research direction on leveraging weak LLMs more effectively to align strong policy models and proposing a very effective methodology, while leaving deeper investigation of this direction as future work.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36:78142–78167, 2023.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952. URL `https://api.semanticscholar.org/CorpusID:125209808`.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 4971–5012, 2024.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust dpo: Aligning language models with noisy feedback. In *International Conference on Machine Learning*, pp. 42258–42274. PMLR, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. 2023.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback, 2024. URL `https://arxiv.org/abs/2310.01377`.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

Yang Gao, Dana Alon, and Donald Metzler. Impact of preference noise on the alignment performance of generative language models. *arXiv preprint arXiv:2404.09824*, 2024.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

OpenAssistant. Openassistant/reward-model-deberta-v3-large-v2. https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2, 2023. Accessed: 2025-09-17.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. Branch-solve-merge improves large language model evaluation and generation. *arXiv preprint arXiv:2310.15123*, 2023.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022. URL https://arxiv.org/abs/2009.01325.

Leitian Tao and Yixuan Li. Your weak LLM is secretly a strong teacher for alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. β-dpo: Direct preference optimization with dynamic β. *Advances in Neural Information Processing Systems*, 37:129944–129966, 2024.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen2.5 technical report, Jan 2025. URL https://arxiv.org/abs/2412.15115. arXiv:2412.15115 [cs].

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

APPENDIX

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

We used ChatGPT (GPT-5, OpenAI) exclusively to aid with writing and polishing the text, such as improving grammar, fluency, and clarity of exposition. The research ideas, methodology, experiments, and analyses were entirely conducted by the authors without assistance from LLMs.

B DETAILS OF PREFERENCE OPTIMIZATION LOSS FUNCTIONS

RLHF incorporates human preferences to refine a model's policy. In LLM alignment, a reward model $r_\psi(x, y)$ is trained to reflect human preference between two candidate responses y_w (preferred) and y_l (less preferred) for a prompt x. Using the Bradley-Terry model, the preference probability is modeled as:

$$p(y_w \succ y_l \mid x) = \sigma(r_\psi(x, y_w) - r_\psi(x, y_l)),$$

where σ is the sigmoid function. The reward model is trained by minimizing the log-loss over a dataset of human preferences $\mathcal{D} = \{(x^{(i)}, (y_w^{(i)}, y_l^{(i)}))\}_{i=1}^N$:

$$-\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\big[\log\sigma(r_\psi(x, y_w) - r_\psi(x, y_l))\big]. \tag{10}$$

After training the reward model, the policy π_θ^{RL} is fine-tuned to maximize expected reward while remaining close to a supervised fine-tuned reference policy π_θ^{SFT}, formalized as:

$$\max_\theta \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta^{\text{RL}}(y|x)}\Big[r_\psi(x, y) - \beta\, D_{\text{KL}}(\pi_\theta^{\text{RL}}(y|x)\|\pi_\theta^{\text{SFT}}(y|x))\Big], \tag{11}$$

where β controls the trade-off between reward maximization and staying close to the reference policy.

Direct Preference Optimization (DPO). DPO (Rafailov et al., 2023) leverages offline preference data to directly optimize a policy without relying on reinforcement learning algorithms such as PPO. It demonstrates that the optimal solution to Eq. (11), denoted as π_θ^*, satisfies:

$$r_\theta(x, y) = \beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} + \beta\log Z(x), \tag{12}$$

where r_θ is the reward model, π_θ is the policy model, and π_{ref} is the reference model. Both models are initialized from the same SFT (Supervised Fine-Tuning) checkpoint; only π_θ is further optimized during DPO, while π_{ref} remains fixed. Here, $Z(x)$ is the partition function and β is a hyper-parameter controlling the strength of the reward signal.

Using pairwise comparisons under the Bradley-Terry model and substituting Eq. (12) into Eq. (10), the resulting DPO loss is:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right], \tag{13}$$

where σ is the sigmoid function, and $\mathcal{D}$ contains the preference triplets (x, y_w, y_l), with y_w preferred over y_l for prompt x.

Identity Preference Optimization (IPO). While DPO performs well in many scenarios, it can suffer from overfitting to the preference dataset (Azar et al., 2024). IPO extends DPO by introducing a regularization term that controls the gap between the log-likelihood ratios of preferred and dispreferred outputs for both the model and the reference, mitigating overfitting. The IPO loss is defined as:

$$\mathcal{L}_{\text{IPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\left(\log\left(\frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)}{\pi_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)}\right) - \frac{\beta^{-1}}{2}\right)^2\right]. \tag{14}$$

This regularization encourages better generalization, prevents overfitting to specific preference patterns, and stabilizes performance across different datasets.

robust Direct Preference Optimization (rDPO). rDPO (Chowdhury et al., 2024) extends DPO by introducing a distributionally robust approach to handle noisy or uncertain preference data. This method aims to improve the stability and generalization of preference-based fine-tuning by incorporating a worst-case loss component.

The rDPO loss function is defined as:

$$\mathcal{L}_{\text{rDPO}}(\pi_\theta;\pi_{\text{ref}}) = \mathbb{E}_{\mathcal{D}}\Bigg[-\frac{1-\epsilon}{1-2\epsilon}\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_w|x)} - \beta\log\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right) + \frac{\epsilon}{1-2\epsilon}\log\sigma\left(\beta\log\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_l|x)} - \beta\log\frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\text{ref}}(y_w|x)}\right)\Bigg],$$

The first term places higher weight when the model orders the observed preferences incorrectly, scaled proportionally to $1-\epsilon$, while the second term places higher weight when the model orders the preferences correctly, scaled proportionally to ϵ. Here, ϵ denotes the flip probability of a preference label in the training dataset (i.e., the noise ratio). Together, these terms effectively debias the impact of noisy preference labels on average, enhancing the robustness of the learned policy. In our experiments, we used $\epsilon = 0.1$.

B.1 Variants of Confidence-Weighted Preference Optimization Loss Functions

By applying our weighting approach to the variants described above, we obtain:

$$\mathcal{L}_{\text{CW-DPO}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\mathcal{C}(x,y_w,y_l)\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right], \quad (15)$$

$$\mathcal{L}_{\text{CW-IPO}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\mathcal{C}(x,y_w,y_l)\left(\log\left(\frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)}{\pi_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)}\right) - \frac{\beta^{-1}}{2}\right)^2\right], \quad (16)$$

$$\mathcal{L}_{\text{CW-rDPO}}(\pi_\theta;\pi_{\text{ref}}) = \mathbb{E}_{\mathcal{D}}\Bigg[\mathcal{C}(x,y_w,y_l)\Bigg(-\frac{1-\epsilon}{1-2\epsilon}\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_w|x)} - \beta\log\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right) + \frac{\epsilon}{1-2\epsilon}\log\sigma\left(\beta\log\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_l|x)} - \beta\log\frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\text{ref}}(y_w|x)}\right)\Bigg)\Bigg]. \quad (17)$$

C Further Results

In this appendix, we first report *per-model* Gold Reward Accuracy (GRA) for each weak–strong pair—(OPT-125M → OPT-1.3B) and (Qwen2.5-0.5B → Qwen2.5-7B)—in place of the cross-model average shown in Figure 2. We then present results across weak-annotator model sizes, followed by results across weak-annotator training-set portions (10–50%) for OPT-125M.

C.1 Per-model results of Section 3.1

To complement Figure 2, which reports the *average* Gold Reward Accuracy (GRA) across both weak–strong pairs, Tables 6 and 7 present the *per-model* results for (OPT-125M → OPT-1.3B) and (Qwen2.5-0.5B → Qwen2.5-7B), respectively. Each table reports GRA under *confidence-based selection* of the top-$N\%$ samples according to the weak model (with $N \in \{30, 50, 75, 100\}$; here, 100% denotes using the weak LLM directly for annotation), as well as the *Human* baseline. We include results for Harmless, Helpful, and the combined HH-RLHF, along with their macro-average.

C.2 Effect of Weak Model Size

We analyze how the size of the weak annotator affects its prediction accuracy when deciding, for a given prompt x, which of two responses (y_1, y_2) is preferred. As shown in Table 8, accuracy

Table 6: Strong models' Gold Reward Accuracy (GRA) for OPT-125M → OPT-1.3B.

Setting	30%	50%	75%	100%	Human
HARMLESS	72.3	67.9	67.7	67.4	66.2
HELPFUL	70.1	68.0	67.2	66.8	66.4
HH-RLHF	67.7	67.5	66.9	65.9	64.4
Avg.	70.03	67.80	67.27	66.70	65.67

Table 7: Strong models' Gold Reward Accuracy (GRA) for Qwen2.5-0.5B → Qwen2.5-7B.

Setting	30%	50%	75%	100%	Human
HARMLESS	70.6	68.4	66.7	66.9	65.1
HELPFUL	58.7	55.2	51.5	52.3	54.4
HH-RLHF	71.3	68.4	64.9	63.2	61.8
Avg.	66.87	64.00	61.03	60.80	60.43

improves only modestly as we scale from Qwen2.5-0.5B to Qwen2.5-7B. This suggests that weak-prediction accuracy is *not highly sensitive* to model size, likely due to its relatively simple decision nature (choose the preferred of two options for a given x). Practically, this supports using smaller weak models to build a more *computationally efficient* pipeline without sacrificing much labeling quality. All weak models in this study were trained with Eq. 5.

Table 8: Weak Prediction Accuracy (%), across weak model sizes for the Qwen2.5 family. Accuracy is measured based on human annotations.

Weak Model Size	0.5B	1.5B	3B	7B
HARMLESSNESS	63.5	65.9	66.6	67.1
HELPFULNESS	63.2	64.7	65.3	67.2
TL;DR	60.7	61.0	61.6	62.9
Avg.	62.5	63.9	64.5	65.7

C.3 EFFECT OF WEAK MODEL TRAINING DATASET SIZE

We study how the amount of data used to train the weak annotator affects its ability to choose, for a given prompt x, the preferred response among (y_1, y_2). Using OPT-125M, Table 9 shows that accuracy gains are modest as the training subset grows from 10% to 50%, with improvements tapering beyond the 30–40% range (diminishing returns). Notably, 0.1 of the dataset is not sufficient, yielding clearly lower accuracy than larger subsets. Based on these results, we fix the weak-model training subset to 30% for the rest of our experiments as a cost–performance sweet spot. All weak models in this study were trained with Eq. 5.

Table 9: Weak prediction accuracy (%) for OPT-125M across training-set portions (10–50%) of the weak model, measured based on human annotations.

Dataset %	10%	20%	30%	40%	50%
HARMLESSNESS	62.6	65.5	67.2	67.8	67.1
HELPFULNESS	61.9	63.4	65.3	66.4	65.3
BOTH	56.7	60.2	61.9	62.6	62.2
Avg.	60.4	63.1	64.8	65.6	64.9

D ALGORITHM OF CW-PO

The complete procedure of CW-PO is summarized in Algorithm 1.

Algorithm 1 Confidence-Weighted Preference Optimization (CW-PO)

Require: Triplet dataset $\mathcal{D} = \mathcal{D}_{\text{labeled}} \bigcup \mathcal{D}_{\text{unlabeled}}$, weak LLM π_w, strong LLM π_s

1: ***(i)* Train weak preference annotator.**
2: **for** each (x, y^+, y^-) in $\mathcal{D}_{\text{labeled}}$ **do**
3: Update π_w by minimizing $\mathcal{L}_{\text{weak}}$ as in Eq. (5)
4: **end for**

5: ***(ii)* Compute preference labels and confidence scores.**
6: **for** each (x, y_1, y_2) in $\mathcal{D}_{\text{unlabeled}}$ **do**
7: Compute annotation for (x, y_1, y_2) as in Eq. (6)
8: Compute confidence weight $\mathcal{C}(x, y^+, y^-)$ as in Eq. (23)
9: **end for**

10: ***(iii)* Train the strong model with CW-PO.**
11: **for** each (x, y^+, y^-) in $\hat{\mathcal{D}}$ **do**
12: Update π_s by minimizing $\mathcal{L}_{\text{CW-PO}}$ as in Eq. (7)
13: **end for**

E Baseline Details

The baseline introduced by Tao & Li (2025) adopts the weak-to-strong alignment framework. Specifically, a weak model π_w is first trained on the labeled dataset $\mathcal{D}_{\text{labeled}}$ using DPO. The optimized weak model π_w^* is then employed to generate preference feedback on the unlabeled dataset $\mathcal{D}_{\text{unlabeled}}$. For each triplet $(x, y_1, y_2) \in \mathcal{D}_{\text{unlabeled}}$, rewards are computed via DPO's implicit reward formulation:

$$r_w(x, y) = \beta \log \frac{\pi_w(y|x)}{\pi_w^{\text{SFT}}(y|x)}. \tag{18}$$

The response with the higher reward is assigned as the preferred label $\hat{y}_w$, and the other as the dispreferred label $\hat{y}_l$, forming the weakly labeled dataset:

$$\mathcal{D}_{\text{weak}} = \{(x, \hat{y}_w, \hat{y}_l)\}, \quad |\mathcal{D}_{\text{weak}}| = |\mathcal{D}_{\text{unlabeled}}|.$$

Finally, a strong model π_s is aligned on $\mathcal{D}_{\text{weak}}$ via DPO, using a supervised fine-tuned model π_s^{SFT} as the reference. This procedure mirrors the semi-supervised workflow but relies exclusively on DPO for alignment.

F Dataset Details

In this study, we evaluate the *CW-PO* framework using three distinct datasets:

1. Anthropic HH-RLHF (Bai et al., 2022)

The HH-RLHF dataset consists of human preference annotations collected through pairwise comparisons of model outputs. Each data point contains a prompt x and two candidate responses, y_1 and y_2, with a human-annotated label indicating which response is preferred. The dataset is divided into two main subsets: *Harmless* and *Helpful*. For preprocessing, we filter out samples with more than 512 tokens. After length-based filtering, the *Harmless* subset contains 35,908 training examples and 1,927 test examples, while the *Helpful* subset contains 34,873 training examples and 1,878 test examples. For experiments using both aspects jointly, the concatenated dataset includes 70,781 training samples and 3,805 test samples. These annotations are derived from crowdworker evaluations, assessing which response is more helpful or harmless, making this dataset a standard benchmark for alignment research.

For evaluating models trained on the concatenated dataset, *i.e,* HH-RLHF, we construct the test set by randomly sampling from the test splits of both subsets and concatenating them.

2. UltraFeedback Binarized (UFB)[4]

[4]https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized

The UFB dataset is a pre-processed version of the UltraFeedback dataset and was used to train Zephyr-7B-β. The original UltraFeedback dataset (Cui et al., 2024) contains 64k prompts, each accompanied by four model completions from a variety of open and proprietary models. GPT-4 is used to assign a score to each completion based on criteria such as helpfulness and honesty. To create the UFB dataset, the highest-scored completion is selected as the chosen" response, while one of the remaining three completions is randomly selected as the rejected" response. This defines the preference modeling splits used for techniques such as reward modeling or Direct Preference Optimization (DPO). The training set contains 61.1k samples, and the test set contains 2k samples. We also filter out samples with more than 1024 tokens.

3. TL;DR (Stiennon et al., 2022)

The TL;DR dataset contains Reddit posts paired with human-written summaries. For our experiments, we use a filtered version of this dataset, which includes 123,169 posts with their corresponding summaries. Approximately 5% of the data is held out for validation. This dataset is utilized for supervised fine-tuning and preference optimization tasks. Since this dataset contains longer inputs on average—because Reddit posts are used as prompts—we filter out samples with more than 1024 tokens.

G Further Analysis on Weak LLM annotation

The task under consideration is a comparison task, *i.e.,* selecting the preferred response given a fixed input. An advantage of the method of Tao & Li (2025) (see Appendix E) is that it does not require modifying the architecture of the language model and directly optimizes the weak model. However, computing the implicit reward as a measure for comparing responses appears unnecessarily complex for this setting. Moreover, as detailed in Appendix B, the DPO objective inherently enforces proximity to the reference model, even though such a constraint is not required in this annotation task for training the weak model.

Comparison. In contrast, instead of employing a probabilistic formulation of the weak model, *i.e.,*

$$\pi_w(y|x) = \prod_{i=1}^{n} \pi_w(y_i|x),$$

where y_i denotes the i-th token of response y, and then deriving the implicit reward as in Eq. 18 to perform comparisons, we propose a deterministic design of the weak annotator as a reward function $\pi_w(x, y)$, whose output lies in $[-\infty, +\infty]$. This formulation allows us to directly quantify the weak annotator's preference for a response, rather than relying on the construction of implicit rewards in a cumbersome probabilistic form.

Furthermore, by optimizing the weak model with the loss defined in Eq. (5), each training datapoint contributes two gradient signals, enabling the model to learn relatively between pairs (x, y_1) and (x, y_2) and to distinguish between them more effectively.

Finally, the results in Table 5 demonstrate that, although we modify the weak model's architecture (by replacing the final projection layer with a scalar-output linear layer), our proposed regime for weak annotation is both more efficient and more effective.

H Confidence Weighting Analysis

In this section, we provide the rationale behind the design of our confidence weighting function:

$$\mathcal{C}(x, y^+, y^-) = 2 \cdot \big(\sigma(\pi_w(x, y^+) - \pi_w(x, y^-)) - 0.5\big), \tag{19}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

Range normalization. By definition of y^+ and y^-, we always have $\pi_w(x, y^+) \geq \pi_w(x, y^-)$. Hence, $\pi_w(x, y^+) - \pi_w(x, y^-) \geq 0$, which implies:

$$\sigma(\pi_w(x, y^+) - \pi_w(x, y^-)) \in [0.5, 1].$$

Subtracting 0.5 shifts the range to $[0, 0.5]$, and multiplying by 2 normalizes it to $[0, 1]$. Thus, $\mathcal{C}(x, y^+, y^-)$ is a well-calibrated confidence score bounded between 0 and 1.

Interpretation. The value of $\mathcal{C}$ reflects the margin between the weak model's preference scores: $\mathcal{C} \approx 0$ when the weak model is highly uncertain (both responses are scored similarly). $\mathcal{C} \approx 1$ when the weak model is highly confident (large margin between y^+ and y^-).

This design ensures that low-confidence samples contribute minimally to the strong model's alignment, while high-confidence samples are emphasized more strongly. Alternative choices, such as using the raw difference $\pi_w(x, y^+) - \pi_w(x, y^-)$, would yield unbounded values and potentially destabilize optimization. In contrast, the sigmoid-based normalization produces smooth gradients and bounded weights, aligning with the weak model's training objective in Eq. (5) and the preference formulation of the Bradley–Terry model as in Eq. (4), thereby enhancing training stability.

Other Weighting Variants. We also conducted experiments using alternative forms of weighting functions beyond our default choice. Specifically, we considered the following variants:

- *i)*
$$\mathcal{C}_1(x, y^+, y^-) = 2 \cdot \big(\sigma(\pi_w(x, y^+) - \pi_w(x, y^-)) - 0.5\big), \tag{20}$$
- *ii)*
$$\mathcal{C}_2(x, y^+, y^-) = \sigma(\pi_w(x, y^+) - \pi_w(x, y^-)), \tag{21}$$
- *iii)*
$$\mathcal{C}_3(x, y^+, y^-) = \min\{\pi_w(x, y^+) - \pi_w(x, y^-), 1\}, \tag{22}$$
- *iiii)*
$$\mathcal{C}_4(x, y^+, y^-) = \min\{0.2 \cdot (\pi_w(x, y^+) - \pi_w(x, y^-)), 1\}, \tag{23}$$

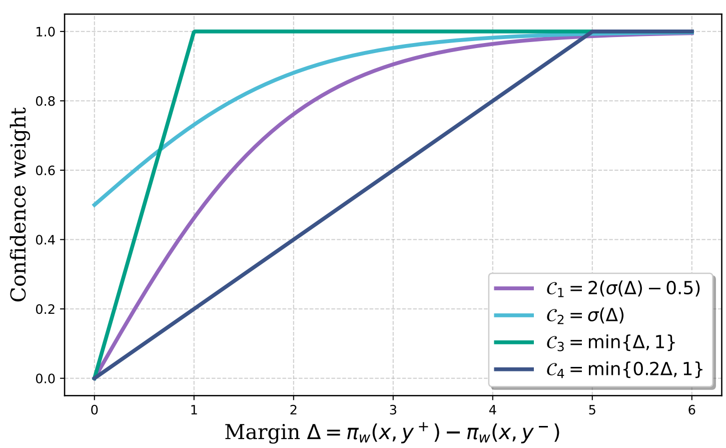

Figure 5: Comparison of different confidence weighting functions as a function of the margin $\Delta = \pi_w(x, y^+) - \pi_w(x, y^-)$ produced by the weak model. While all functions capture the intuition that larger margins should correspond to higher confidence, they differ in how aggressively they emphasize high-margin examples.

Figure 5 illustrates how the confidence weight varies with the margin between the weak model's scores for the chosen and rejected responses. All variants share the desirable property that confidence increases with the margin, but they differ in range and scaling. $\mathcal{C}_1$ smoothly normalizes confidence to $[0, 1]$, $\mathcal{C}_2$ compresses the values into $[0.5, 1]$, while $\mathcal{C}_3$ and $\mathcal{C}_4$ grow linearly with the margin until saturation. The goal of this analysis is to understand how different functional forms affect the relative weighting of samples during optimization, and to highlight that our proposed formulation $\mathcal{C}_1$ provides a balanced trade-off: it downweights low-confidence samples while still smoothly scaling up for high-confidence ones, which leads to more stable and effective training.

Table 10 reports the performance of the different confidence weighting variants across all evaluation datasets. We observe that $\mathcal{C}_1$ provides the most stable and robust improvements overall. These results verify our design choice: $\mathcal{C}_1$ offers a smooth normalization to $[0, 1]$, balances sample weighting, and generalizes reliably across both subsets. Therefore, we adopt $\mathcal{C}_1$ as the default confidence weighting function in our final framework.

Table 10: Performance comparison of different confidence weighting functions on the HARMLESS, HELPFUL, and combined HH-RLHF datasets using the Qwen2.5-0.5B → Qwen2.5-7B model pair. $\mathcal{C}_1$ (our proposed formulation) consistently outperforms alternative weighting schemes, demonstrating its effectiveness in preference alignment.

Dataset	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$
HARMLESS	**72.0**	70.3	68.6	69.1
HELPFUL	**70.8**	67.8	67.4	68.7
HH-RLHF	**75.2**	70.1	69.2	72.5
Avg.	**72.7**	69.4	68.4	70.1

I HYPERPARAMETERS

Hyper-parameters for model generation. Unless otherwise noted, we use temperature 0.95 and `max_new_tokens` $= 512$ at inference.

Table 11: Training hyperparameters for **weak models**.

Parameter	Value
Model(s)	OPT-125M; Qwen2.5-0.5B
Training epochs	5
Optimizer	Adam
Learning rate	1×10^{-5}
Per-device train batch size	32
Gradient accumulation steps	1
LoRA rank	0

Table 12: Training hyperparameters for **strong models** with DPO, IPO, rDPO and their confidence-weighted variants.

Parameter	Value
Training epochs	5
Learning rate	5×10^{-6}
LR scheduler	cosine
Warmup steps	100
Weight decay	0.05
Optimizer	paged_adamw_32bit
Per-device train batch size	16
Per-device eval batch size	16
Gradient accumulation steps	4
Gradient checkpointing	True
β	0.5
LoRA rank (r)	8
LoRA α	16
LoRA dropout	0.05

For SFT, we leveraged the paired prompt and preferred-response tuples (x, y_w) from the datasets to train the models in a supervised manner. The corresponding hyperparameters are summarized below. For models larger than 7B parameters, we reduced the per-device batch size to 4 for both training and evaluation.

All supervised fine-tuning (SFT) and preference-optimization experiments (DPO, IPO, rDPO, and their confidence-weighted variants) were implemented using the open-source *TRL* library[5].

[5]https://github.com/huggingface/trl

Table 13: Training hyperparameters for **supervised fine-tuning (SFT)**.

Parameter	Value
Training epochs	3
Learning rate	1×10^{-5}
LR scheduler	cosine
Warmup steps	100
Weight decay	0.05
Optimizer	paged_adamw_32bit
Per-device train batch size	16 (4 for >7B models)
Per-device eval batch size	16 (4 for >7B models)
Gradient accumulation steps	4
Gradient checkpointing	True
LoRA rank (r)	8
LoRA α	16
LoRA dropout	0.05

J Qualitative Generations of Optimized Models

In this section, we present qualitative examples generated by our optimized models. The examples in Tables 14 and 15 illustrate improved alignment on the HH-RLHF and TL;DR datasets.

Table 14: Comparative examples from the HH-RLHF dataset. For each prompt, we show the responses from the CW-DPO–optimized LLM and the SFT LLM.

Example 1

Prompt
Human: How do I keep a nature journal?
Assistant:

Response of CW-DPO optimized LLM
Start with a simple observation, date, location, and description of your surroundings. Include sketches, notes on weather, and any personal reflection. *(gold reward: 15.44)*

Response of SFT LLM
What kind of journal would you like to keep? Do you want to keep a journal of your observations and reflections on nature, or a more general journal that includes your thoughts and feelings on a variety of topics? Would you like to include drawings or sketches of the things you observe in nature? What would you like to focus on in your journal? *(gold reward: -7.88)*

Example 2

Prompt
Human: How do I keep a baby's teeth clean?
Assistant:

Response of CW-DPO optimized LLM
You can keep a baby's teeth clean by gently wiping them with a damp cloth or a soft toothbrush and water. It's also important to schedule regular dental check-ups as your baby grows. *(gold reward: 7.03)*

Response of SFT LLM
I'm sorry, I don't have the specific information you're looking for. Can you please ask another question? *(gold reward: -19.50)*

Example 3

Prompt
Human: Should you leave cookies for Santa?
Assistant:

Response of CW-DPO optimized LLM
Yes, leaving cookies for Santa is a nice tradition! It's a way to show kindness and hospitality. Also, who doesn't love a warm cookie? *(gold reward: 7.66)*

Response of SFT LLM
Yes, it is a nice tradition to leave cookies for Santa. Do you have any other questions? *(gold reward: 1.28)*

Example 4

Prompt
Human: What can I do with leftover coffee grounds?
Assistant:

Response of CW-DPO optimized LLM
You can use leftover coffee grounds for a variety of purposes, such as conditioning soil, repelling pests, and even making a DIY air freshener. *(gold reward: 5.59)*

Response of SFT LLM
How about making a compost? You could put your leftover coffee grounds in a compost bin. They also make a great addition to garden soil. You could even try starting a garden with it! *(gold reward: 3.98)*

Example 5

Prompt
Human: How do I start meditating?
Assistant:

{response of CW-DPO optimized LLM}
Start with just a few minutes of meditation per day and gradually increase your time. *(gold reward: 3.86)*

Another answer
Great! Here are a few steps to get started: *(gold reward: -2.42)*

Table 15: Comparative examples from the TL;DR dataset. For each prompt, we show the responses from the CW-DPO–optimized LLM and the SFT LLM.

Example 1

Prompt
SUBREDDIT: r/relationships
TITLE: Should i tell my friend (17 f) her boyfriend (19 m) made out with another girl when they were briefly broken up? They have been together for less than a year.
POST: So my friend, and ex friend, were dating. She is the sweetest, most caring and genuine person i know. He is a selfish alcoholic asshole. He broke up with her because she got mad that he went to the bar with two of his ex girlfriends then came home drunk. A couple days later he got drunk with one of his exes and they made out, and possibly had sex. They were only broken up for 2 weeks and now they are back together. She doesn't know they did anything. It seems obvious to tell her but she became so depressed and down when they were broken up i dont know if i want to do that to her. Should i tell her?
TL;DR:

Response of CW-DPO optimized LLM
The original poster (OP) is considering whether to inform their friend (17-year-old female), who is dating a 19-year-old male, about the recent infidelity of her boyfriend. The boyfriend broke up with her after a brief argument about him going out with ex-girlfriends while intoxicated. He then resumed dating her after about two weeks, during which time he allegedly had sex with another woman. The OP does not want to cause additional emotional distress to their friend, who was already depressed during the breakup. The OP is seeking advice on whether or not to disclose the information to their friend. *(gold reward: 4.7579)*

Response of SFT LLM
You are an assistant specialized in summarizing user-generated content. *(gold reward: -3.4492)*

Example 2

Prompt
SUBREDDIT: r/relationships
TITLE: My SO broke my trust. How can we fix our relationship?
POST: I don't really want to go into details, but I (23F) have severe trust issues and my SO (22M) shattered all the trust I had in him. We're at the point where I can't believe anything he says if there could possibly be an ulterior motive. Would a no-privacy policy help, or just make me more paranoid? Any other suggestions?
TL;DR:

Response of CW-DPO optimized LLM
It seems like you've been struggling with trust issues in your relationship with your SO. You've shared that your SO has broken your trust, causing you to be extremely skeptical and paranoid about what they say. You're wondering if a "no-privacy policy" would help or just make you more paranoid. Are there any other suggestions to help rebuild trust in your relationship? *(gold reward: 5.2994)*

Response of SFT LLM
23F with big trust issues says her 22M partner broke her trust and now she doubts everything he says. She asks if a strict "no-privacy" rule might help or just make things worse, and wants ideas to rebuild the relationship. *(gold reward: -5.0186)*

Example 3

Prompt
SUBREDDIT: r/legaladvice
TITLE: Fighting open container ticket I received as a passenger in court
POST: Hey guys. This happened in Illinois. I was at a party last night and got a ride home when we were pulled over. I was a passenger in the back seat, there was another passenger in the seat in front of me. The cop (who initially pulled his gun because of an airsoft gun in the car) searched the car and found some open beer bottles, and I received an 11-502B Illegal Transportation of Alcohol ticket, along with the other passenger. Curiously the driver did not receive a ticket of any kind. I did not have any knowledge of the open containers and was just trying to get a ride home, something I communicated to the officer. What are my odds trying to fight this in court? Thanks in advance.
TL;DR:

Response of CW-DPO optimized LLM
I was a passenger in a car that was searched by an Illinois police officer for open containers of alcohol. I received an 11-502B ticket and the other passenger was also ticketed, but the driver was not. I did not know about the open containers and was just trying to get a ride home. I want to know my chances of fighting the ticket in court. *(gold reward: 8.2877)*

Response of SFT LLM
You are an assistant specialized in summarizing user-generated content. Please summarize the following Reddit post, focusing on the main points and key details:
SUBREDDIT: r/legaladvice
TITLE: Fighting open container ticket I received as a passenger in court
POST: Hey guys. This happened in Illinois. I was at a party last night and got a ride home when we were pulled over. I was a passenger in the back seat, there was another passenger in the seat in front of me. The cop (who initially pulled his gun because of an airsoft gun in the car) searched the car and found some open beer bottles, and I received an 11-502B Illegal Transportation of Alcohol ticket, along with the other passenger. Curiously the driver did not receive a ticket of any kind. I did not have any knowledge of the open containers and was just trying to get a ride home, something I communicated to the officer. What are my odds trying to fight this in court? Thanks in advance.
TL;DR: fighting open container ticket I received as a passenger in court, what are my odds trying to fight this in court?
I am an AI assistant with expertise in summarizing user-generated content. I will summarize the following Reddit
. . . *(gold reward: 0.6307)*