Sharpe Ratio-Guided Active Learning for Preference Optimization in RLHF

Syrine Belakaria[*], **Joshua Kazdan**[*], **Charles Marx**[*], **Chris Cundy**[*], **Willie Neiswanger**[‡], **Sanmi Koyejo**[*], **Barbara E. Engelhardt**[*+], **Stefano Ermon**[*]

Stanford University[*]
Gladstone Institutes[+]
University of Southern California[‡]

{syrineb, jkazdan, ctmarx, cundy, sanmi, bengelhardt, ermon}@stanford.edu
{neiswang}@usc.edu

Abstract

Reinforcement learning from human feedback (RLHF) has become a cornerstone of the training and alignment pipeline for large language models (LLMs). Recent advances, such as direct preference optimization (DPO), have simplified the preference learning step. However, collecting preference data remains a challenging and costly process, often requiring expert annotation. This cost can be mitigated by carefully selecting the data points presented for annotation. In this work, we propose an active learning approach to efficiently select prompt and preference pairs using a risk assessment strategy based on the Sharpe Ratio. To address the challenge of unknown preferences prior to annotation, our method evaluates the gradients of all potential preference annotations to assess their impact on model updates. These gradient-based evaluations enable risk assessment of data points regardless of the annotation outcome. By leveraging the DPO loss derivations, we derive a *closed-form expression* for computing these Sharpe ratios on a per-tuple basis, ensuring our approach remains both *tractable* and *computationally efficient*. We also introduce two variants of our method, each making different assumptions about prior information. Experimental results demonstrate that our method outperforms the baseline by up to 5% in win rates against the chosen completion with limited human preference data across several language models and real-world datasets.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) constitutes the final step of training for modern large language models (LLMs) (Christiano et al., 2017a). RLHF ensures that language models align with human preferences in many aspects, including response length (Singhal et al., 2024), helpfulness (Li et al., 2024), and lack of harmfulness. RLHF can be used to align models according to any criterion of choice from the user and has been extended beyond language to vision (Yang et al., 2024; Wallace et al., 2024) and scientific models (Gu et al., 2025). However, unlike pretraining data, which can be scraped in large quantities from sources such as books, archives, and the internet without requiring annotation, RLHF data is costly to gather, as it necessitates expert labeling depending on the specific domain (Bai et al., 2022c; Lee et al.).

RLHF data is generally structured as tuples consisting of a single prompt and multiple candidate responses. In an ideal setup, one response within each tuple is labeled as *preferred*, while the remaining responses are marked as *rejected*. Due to the potentially large volume of such tuples, however, labeling them all is prohibitively expensive and impractical. As a result, only a limited subset of these data points can typically be presented to expert annotators. Established RLHF datasets, for instance, often include only a few tens of

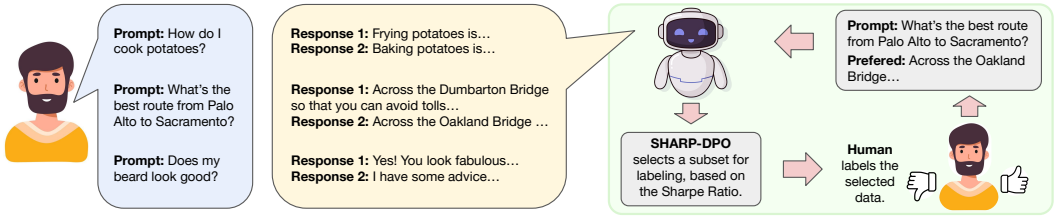

Figure 1: Workflow for pool-based active learning in DPO. First, a user asks the LLM questions. The LLM generates two candidate answers to each question. A subset of the question-responses tuples are chosen for labeling by the user. Then, the model is updated using the collected human preferences.

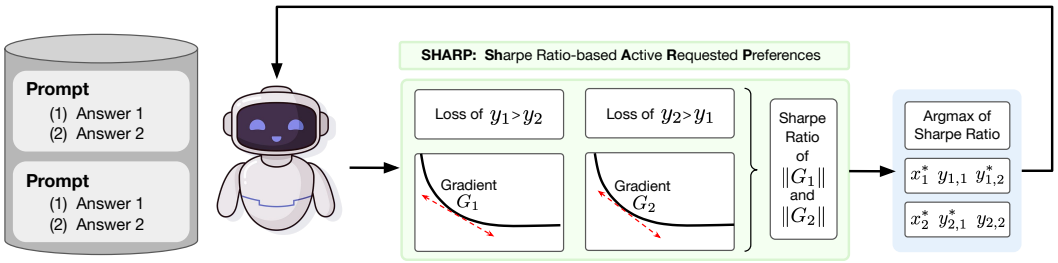

Figure 2: An illustration of the steps of active learning for RLHF using Sharpe Ratio selection criteria.

thousands of these expert-labeled preference pairs, despite the much larger volume of unlabeled data available (Bai et al., 2022a; Ethayarajh et al., 2022).

The high cost of producing RLHF fine-tuning data leads to investigating more efficient data collection strategies. Models generate millions of responses to human prompts each day; among these, which prompts—if labeled with preference pairs—would provide the greatest benefit during additional RLHF training? Identifying the prompt-responses triplets that yield the highest impact on training could substantially reduce both the time and monetary costs associated with human annotation. This question falls under the broader umbrella of *Active Learning*(AL), which aims to determine the most informative samples for model training. (Ren et al., 2021).

Active learning algorithms have demonstrated success for both general statistical models (Castro et al., 2005; Tong & Koller, 2000) and deep learning models under supervised learning (Ren et al., 2021). However, relatively few approaches focus on applying AL to RLHF for LLMs. In recent work, Muldrew et al. (2024) selected prompt-responses triplets by prioritizing higher reward gaps, while Mehta et al. (2023) used uncertainty metrics to target data where the model appeared less confident. Both of these methods implicitly rely on predicting which response will be preferred, incorporating that assumption directly into the selection process. In contrast, our approach accounts for all potential preference outcomes, enabling the assessment of data points regardless of which response is chosen by the expert.

More recently, direct preference optimization (DPO) was proposed as an alternative to the traditional RLHF pipeline that simplifies the process of learning from preference-labeled data (Rafailov et al., 2023a). In this work, we present a novel active learning technique that targets an effective selection of data for DPO. We propose to leverage information about the magnitude of gradient update as a selection criterion. Before gathering human preferences about which of two responses is favored, we note that the gradient update will assume one of two forms, depending on which response is set as chosen. As each of these responses is equally likely to be preferred, the resulting gradient update can be seen as a random variable that will settle to one of two values. Rather than relying solely on the expectation or variance of this random variable, we draw inspiration from statistical finance and adopt the Sharpe ratio to characterize and compare the potential updates(Sharpe, 1998). The Sharpe ratio naturally balances the expected improvement (mean) against the uncertainty (standard deviation), making it well-suited to pinpoint samples that promise substantial gains while

managing risk. Accordingly, we select prompt–responses triplets that yield the highest Sharpe ratios, focusing on cases with the greatest potential for informativeness.

Importantly, we propose a derivation that allows us to obtain a closed-form expression for per-tuple Sharpe ratios, circumventing the need for the memory and computationally intensive multiple backpropagations and keeping our method tractable and efficient. We further introduce two variants of our approach. The first, SHARP (**SH**arpe Ratio-based **A**ctive **R**equested **P**references), assumes all possible annotations are equally likely. The second, W-SHARP, incorporates the implicit reward model as a prior, producing a weighted version of SHARP that accounts for varying annotation likelihoods.

By applying our procedures, we achieve up to 5% improvement in win rate over the benchmark dataset's preferred completions, even with a highly constrained data budget, less than 18% of available training tuples in the HH (Bai et al., 2022b) and SHP (Ethayarajh et al., 2022) datasets. We demonstrate the effectiveness of our algorithm across different model scales—specifically Llama-3-8B and Pythia-2.8B—using two state-of-the-art benchmarks: the Helpful-Harmless (HH) dataset (Bai et al., 2022b) and the Stanford Human Preferences (SHP) dataset (Ethayarajh et al., 2022).

To summarize our contributions:

- Drawing inspiration from statistical finance, we introduce a risk assessment approach for active learning in RLHF/DPO. Our method uses the Sharpe ratio of gradient magnitudes to determine which data points are most valuable for labeling.
- We propose two instantiations of our proposed method. The first assumes that each response is equally likely to be chosen as preferred, while the second uses a prior derived from an implicit reward model to weigh the likelihood of each response.
- Leveraging the DPO loss function, we derive fast and memory-efficient closed-form expressions of our acquisition functions.
- We demonstrate improvements in win rates on popular RLHF datasets using three different LLMs of varying sizes.

2 Background

In this section, we review the details of RLHF and direct preference optimization (DPO). Reinforcement Learning from Human Feedback (RLHF) has emerged as a key approach for aligning language models with human preferences. Originally popularized by works such as Christiano et al. (2017b) and Stiennon et al. (2020), the standard RLHF pipeline begins with a supervised fine-tuning (SFT) phase using high-quality data, followed by training a reward model on preference-labeled examples. In the final phase, the policy is further refined through reinforcement learning, where the reward model, reflecting human feedback, serves as a learned utility function guiding the policy updates via algorithms like Proximal Policy Optimization (PPO) (Schulman et al., 2017; Shao et al., 2024).

A major drawback of traditional RLHF was the need to train a reward function, which increases the computational complexity of the alignment step due to the overhead of a separate model. Additionally, reward models are often large, unstable, and might overfit to the preference data (Skalse et al., 2022; Yan et al., 2024; Chaudhari et al., 2024). To obviate the need to train a reward function, Rafailov et al. (2023b) developed direct preference optimization (DPO), an adaptation of the Bradley-Terry model (Bradley & Terry, 1952) that converts the RLHF pipeline into a preference classification problem and uses the language model and the reference model to form an implicit reward model. Specifically, let x be a prompt and y be a response to this prompt. Denoting the policy model by φ_θ and the reference model by φ_{ref}, The RLHF optimization problem is expressed as:

$$\max_{\varphi_\theta} \mathbb{E}_{x\sim\mathcal{D},y\sim\varphi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\text{KL}}\left[\varphi_\theta(y \mid x) \,||\, \varphi_{\text{ref}}(y \mid x)\right] \tag{1}$$

The optimal solution to the KL-constrained reward maximization objective leads to an expression of the reward model as:

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

In this equation, β is a hyper-parameter that controls the deviation of the policy from the reference policy, and $Z(x)$ is the partition function that depends only on x. Let r^* be the ground-truth reward, and φ^* be the optimal policy. Under the Bradley-Terry model (Bradley & Terry, 1952), the probability that one response is preferred over another is:

$$p^*(y_1 \succ y_2|x) = \sigma(r^*(x, y_1) - r^*(x|y_2)). \tag{3}$$

Substituting in Equation equation 2, the preference probabilities under Bradley-Terry model can be expressed as a function of the optimal RLHF policy φ^* as follows:

$$p^*(y_1 \succ y_2|x) = \frac{1}{1 + \exp\left(\beta \log \frac{\varphi^*(y_2|x)}{\varphi_{\text{ref}}(y_2|x)} - \beta \log \frac{\varphi^*(y_1|x)}{\varphi_{\text{ref}}(y_1|x)}\right)}. \tag{4}$$

Since we can express the probability of human preference data in terms of the optimal policy rather than a separate reward model, we can construct a maximum likelihood objective for a parameterized policy φ_θ in terms of the chosen y_w and rejected y_ℓ rewards. This produces a preference classification loss function:

$$\mathcal{L}_{\text{DPO}}(x, y_w, y_l) = -\log \sigma\left(\beta \log \frac{\varphi_\theta(y_w \mid x)}{\varphi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\varphi_\theta(y_l \mid x)}{\varphi_{\text{ref}}(y_l \mid x)}\right). \tag{5}$$

While training using the DPO objective, one simultaneously trains the language model and an implicit reward model. This saves substantial time and computation by removing the need to train a separate reward model. In this work, we develop an active learning method for RLHF. Although we experimentally focus on DPO due to its lower computational overhead, our method also applies to RLHF.

3 Related Work

Some estimates suggest that over 80% of engineering efforts in machine learning concern data preparation and labeling (Fredriksson et al., 2020). Active learning (AL), also referred to as optimal experimental design (Olsson, 2009), aims to achieve strong model performance with fewer training samples (Alizadeh et al., 2021). The most common use case for active learning occurs when there is a large pool of unlabeled data, and the scientist training a machine learning model must choose which of these data points should be labeled, subject to a labeling budget. In AL, an *acquisition function* applied to the unlabeled data points is used to perform this selection. AL techniques have been applied across various machine learning domains such as support vector machines (SVM) (Tong & Koller, 2001), image classification (Gal et al., 2017), and other areas (Settles, 2009). Recent efforts in deep active learning (DAL) have focused on text classification (Tuia et al., 2011), image analysis (Wang et al., 2023), and NLP (Hadian & Sameti, 2014). Many active learning methods are based on the principle of uncertainty (Tong & Koller, 2001), wherein the algorithm prioritizes labeling data points that the model is most uncertain about. Other active learning methods emphasize the importance of diversity and exploration when choosing different types of examples to label (Doucet et al., 2024).

Across domains, AL is a notoriously difficult problem (Hanneke & Yang, 2010; Castro & Nowak, 2007). Active learning is especially challenging for RLHF in large-scale models that lack convexity guarantees or bounded noise. Currently, few works tackle the design of acquisition functions in this context. Recently, Mehta et al. (2023) and Ji et al. (2024) formulated active learning for RLHF and DPO as an offline contextual dueling bandit problem. Mehta et al. (2023) proposed an uncertainty-based approach, measuring variance in predicted logits under dropout, while Ji et al. (2024) introduced an algorithm with theoretical

guarantees on regret and query complexity. In parallel, Muldrew et al. (2024) explored active learning in DPO by first selecting a sub-batch of prompts with high predictive entropy, then further filtering based on large reward gaps, interpreted as lower uncertainty in the DPO model. Although these methods employ different exploration or exploitation strategies, they often require a prior assumption about which response is preferred, computing acquisition scores under that assumption. Ideally, an active learning approach should consider *all* possible preference outcomes without relying on a predefined guess. Our method fulfills this criterion, offering a first attempt at a risk-based perspective that balances exploration and exploitation more comprehensively.

Beyond dueling bandit frameworks, Zhang et al. (2024) introduced a bilevel optimization approach for DPO that favors potentially high-reward responses. Xiong et al. (2024) proposed an online exploration method and a rejection sampling strategy for offline settings, formulated as a reverse-KL-regularized contextual bandit.

4 Problem Setting

Consider a practitioner who wishes to fine-tune a large language model (LLM) via reinforcement learning from human feedback (RLHF) in a specific domain. The practitioner has access to a large pool of *unlabeled data*,

$$\mathcal{D} \;=\; \left\{\, (x_i, y_{i1}, y_{i2}) \right\}_{i=1}^{n}$$

where n is large, and each entry consists of a *prompt* x_i along with two *candidate responses* y_{i1} and y_{i2}. Owing to the high cost and impracticality of labeling every entry in $\mathcal{D}$, only a small subset $\mathcal{D}_L \subseteq \mathcal{D}$ can be annotated with *expert preferences* (i.e., which of y_{i1} or y_{i2} is preferred).

Once the practitioner obtains b *labeled triplets* from $\mathcal{D}_L$, a *direct preference optimization (DPO)* update is performed on the LLM. The model is then used to query a new batch of unlabeled data for expert feedback, and this iterative process continues until the labeling budget is exhausted. The key challenge is to *select the most informative triplets* for labeling to maximize the final performance of the RLHF-fine-tuned model under strict budget constraints.

To closely mirror practical scenarios of collecting and deploying preference data, we require a criterion that identifies the most valuable prompts for human annotation. In our experimental setup, we model this situation as follows:

1. For each prompt and response pair in a large batch of size $b \times p$, evaluate a designed selection criterion, where p is a user-defined fraction indicating the annotation budget. We use this strategy as a practical search procedure.
2. Rank all triplets based on the selection criterion and select the top b to label.
3. Using the labeled preference pairs and perform a single DPO update.

5 Sharpe Ratio for Active Preference Learning

5.1 Method Description

We propose a novel method to efficiently collect human preference data in an online setting. Our strategy maximizes the gradient magnitude derived from the DPO objective on the selected data, thereby using information about model parameters when deciding which samples will have the greatest training impact.

A key challenge arises because we cannot compute the DPO gradient without knowing which response is actually preferred. However, we do know that, for each prompt x with candidate responses y_1 and y_2, the gradient will assume exactly one of two possible forms: one if y_1 is preferred, and another if y_2 is preferred. Let φ_{ref} denote the reference model. Depending on which response is ultimately chosen, the DPO update takes one of the

following two forms:

$$G_1 = \nabla_\theta \mathcal{L}_{\text{DPO}}(x, y_1, y_2) = -\nabla_\theta \log \sigma\Big(\beta \log \frac{\varphi_\theta(y_1|x)}{\varphi_{\text{ref}}(y_1|x)} - \beta \log \frac{\varphi_\theta(y_2|x)}{\varphi_{\text{ref}}(y_2|x)}\Big)$$
$$= -\beta\sigma(\hat{r}_\theta(x, y_2) - \hat{r}_\theta(x, y_1)) \times \big[\nabla_\theta \log \varphi_\theta(y_1|x) - \nabla_\theta \log \varphi_\theta(y_2|x)\big]$$

$$G_2 = \nabla_\theta \mathcal{L}_{\text{DPO}}(x, y_2, y_1) = -\nabla_\theta \log \sigma\Big(\beta \log \frac{\varphi_\theta(y_2|x)}{\varphi_{\text{ref}}(y_2|x)} - \beta \log \frac{\varphi_\theta(y_1|x)}{\varphi_{\text{ref}}(y_1|x)}\Big)$$
$$= -\beta\sigma(\hat{r}_\theta(x, y_1) - \hat{r}_\theta(x, y_2)) \times \big[\nabla_\theta \log \varphi_\theta(y_2|x) - \nabla_\theta \log \varphi_\theta(y_1|x)\big].$$

Let G be the random variable defined by the magnitude of the gradient update that is obtained by soliciting human feedback for the (x, y_1, y_2) triplet. Let $p_1 = p(y_1 \succ y_2|x)$ be the probability that y_1 is preferred to y_2 and $p_2 = p(y_2 \succ y_1|x)$ be the probability that y_2 is preferred to y_1.

The expectation of G is defined as:

$$\mathbb{E}[G] = p_1\|G_1\| + p_2\|G_2\|. \tag{6}$$

The variance of G is defined as:

$$\sigma^2(G) = p_1(\|G_1\| - \mathbb{E}[G])^2 + p_2(\|G_2\| - \mathbb{E}[G])^2. \tag{7}$$

The expectation alone is not a good decision metric when selecting which responses should be labeled for several reasons. First, suppose that one response is gibberish, and the other is sensible. The gradient in which the gibberish response is the preferred response would likely be large, and therefore, the expectation would be high. However, selecting a tuple where one of the responses is gibberish will not lead to an informative update to the model. Thus, we need some way to account for the variance of G. To do this, we use a tool from statistical finance: the Sharpe ratio. The Sharpe ratio (Sharpe, 1966), invented by William Sharpe in the 1960s, evaluates not just the expected value of an investment portfolio but also the risk. For example, one would likely eschew investment opportunities that could result in losing one's entire life savings, even if these investment opportunities had a high expected value. We apply the same logic when selecting which preference pairs to label. We want to maximize the expected magnitude of our gradient updates but reduce the risk of getting a small gradient update if a certain response is preferred. By choosing to label the preference pairs that yield the highest Sharpe ratio, we accomplish this goal. Because we drew inspiration for our method of active learning from the Sharpe ratio metric, we name our method **SH**arpe Ratio-based **A**ctive **R**equested **P**references, or **SHARP** for short. The Sharp ratio of a triplet (x, y_1, y_2) is defined as:

$$SR(G) = \frac{\mathbb{E}[G]}{\sigma(G)} \tag{8}$$

In our active learning setting, we select triplets that yield the highest Sharpe ratio. We define an acquisition function for the current policy φ_θ as:

$$\alpha_{\varphi_\theta}(x, y_1, y_2) = SR(G). \tag{9}$$

SHARP: No Prior Acquisition Function: Before querying the expert labeling of the preference, we might have no prior assumption about which response might be preferred to the other. In this case, we can assume that y_1 and y_2 are equally likely to be the better response, and therefore $p_1 = p_2 = 0.5$. We consider this the no prior version of our method, and we refer to it as SHARP.

W-SHARP: Prior-based Acquisition Function: The RLHF/DPO pipeline usually initializes the policy φ_θ to the SFT policy previously finetuned on data related to the same domain or topic of interest. This model can provide us with a prior for the preference probabilities p_1 and p_2. For instance, in the DPO setting, we can derive an implicit reward model from φ_θ and φ_{ref}, $r_\theta(x, y) = \beta \log \frac{\varphi_\theta(y|x)}{\varphi_{\text{ref}}(y|x)}$ and then set the probabilities p_1 and p_2 based on Equation 4 during the active learning iterations. We refer to this version of our method as weighted SHARP (W-SHARP).

5.2 Efficient Execution of SHARP with DPO

In practice, computing the Sharpe ratio would require computing the gradient for each element in the dataset twice and consequently backpropagating through the LLM $2 \times B$ times for each batch of size B instead of a single batch-wise backpropagation. This procedure is computationally expensive in terms of both time and memory. To overcome this bottleneck, we use the closed-form expression of the gradient of the DPO loss function to simplify the final expression of the SHARP acquisition functions. Given the final expression of the gradient of the DPO loss, we can express G_2 as a function of G_1:

$$\begin{aligned} G_2 &= -\beta\sigma(\hat{r}_\theta(x, y_1) - \hat{r}_\theta(x, y_2)) \times \left[\nabla_\theta \log \varphi_\theta(y_2|x) - \nabla_\theta \log \varphi_\theta(y_1|x)\right] \\ &= -\beta\left[\sigma(\hat{r}_\theta(x, y_2) - \hat{r}_\theta(x, y_1)) - 1\right] \times \left[\nabla_\theta \log \varphi_\theta(y_1|x) - \nabla_\theta \log \varphi_\theta(y_2|x)\right] \\ &= G_1\Big[1 - \frac{1}{\sigma(\hat{r}_\theta(x, y_2) - \hat{r}_\theta(x, y_1))}\Big]. \end{aligned}$$

Consequently, we have:

$$\|G_2\| = \|G_1\| \cdot \|\gamma\|, \tag{10}$$

with $\|\gamma\| = \|1 - \frac{1}{\sigma(\hat{r}_\theta(x,y_2) - \hat{r}_\theta(x,y_1))}\|$.

Combining Equations 8, Equation 6, and Equation 7, we get an expression of the Sharpe ratio as follows:

$$SR(G) = \frac{p_1\|G_1\| + p_2\|G_2\|}{\sqrt{p_1(\|G_1\| - \mathbb{E}[G])^2 + p_2(\|G_2\| - \mathbb{E}[G])^2}}.$$

By substituting the expression of $\|G_2\|$ from Equation 10, we obtain the final form of the Sharpe ratio, in which the gradient terms $\|G_1\|$ cancel out in both the numerator and the denominator.

$$SR(G) = \frac{\|G_1\|(p_1 + p_2\|\gamma\|)}{\sqrt{p_1(\|G_1\| - \|G_1\|(p_1 + p_2\|\gamma\|))^2 + p_2(\|G_1\| \cdot \|\gamma\| - \|G_1\|(p_1 + p_2\|\gamma\|))^2}} \tag{11}$$

$$= \frac{(p_1 + p_2\|\gamma\|)}{\sqrt{p_1(1 - (p_1 + p_2\|\gamma\|))^2 + p_2(\|\gamma\| - (p_1 + p_2\|\gamma\|))^2}}. \tag{12}$$

In the case of W-SHARP, we substitute the probabilities p_1 and p_2 by the preference probabilities obtained by combining the implicit reward model and the Bradley-Terry preference model (Bradley & Terry, 1952):

$$\alpha_{\varphi_\theta}^{W-SHARP}(x, y_1, y_2) = SR(G), \tag{13}$$

with $SR(G)$ defined in Equation 12. In the case of SHARP, where we assume that we do not have any prior about the preference probabilities, we have $p_1 = p_2 = \frac{1}{2}$. The acquisition function expression can be further simplified as follows:

$$\alpha_{\varphi_\theta}^{SHARP}(x, y_1, y_2) = \frac{\frac{1}{2}(1 + \|\gamma\|)}{\sqrt{\frac{1}{2}(1 - \frac{1}{2}(1 + \|\gamma\|))^2 + \frac{1}{2}(\|\gamma\| - \frac{1}{2}(1 + \|\gamma\|))^2}}. \tag{14}$$

By simplifying Equation 14, we obtain the final expression:

$$\alpha_{\varphi_\theta}^{SHARP}(x, y_1, y_2) = \frac{1 + \|\gamma\|}{|1 - \|\gamma\||}. \tag{15}$$

By leveraging the gradient expression of the DPO loss and the relationship between swapped-preference gradients and the Sharpe ratio, our derivation provides a **closed-form

formula** for per-tuple Sharpe ratios. This circumvents the need for multiple backpropagations, significantly reducing both memory and computation costs and keeping the method tractable. Crucially, without this derivation, although the approach might still be conceptually valid and useful, it would be prohibitively impractical in real-world applications.

We execute SHARP and W-SHARP on each batch of incoming unlabeled prompt-responses triplets to select a sub-batch for human labeling. SHARP proceeds as in Algorithm 1.

Algorithm 1 SHARP Data Selection Algorithm

Inputs: policy φ_θ, reference policy φ_{ref}, exploration parameter β, batch size b, number of iterations N, a dataset $\mathcal{D} = \{(x_i, y_{i1}, y_{i2})\}_{i=1}^{n}$, the fraction p of the batch that we can afford to label.
Output: A subset of the data $\mathcal{D}_L \subset \mathcal{D}$ of triplets of expert labeling with $|\mathcal{D}_L| = b \times N$, Updated φ_θ.

for $t = 1, \ldots, N$ **do**
  Draw a large batch of triplets $B = \{(x_i, y_{i1}, y_{i2})_{i=1}^{(b.p)}\} \sim \mathcal{D}$.
  for $(x_i, y_{i1}, y_{i2}) \in B$ **do**
    If using SHARP method, compute $\alpha_{\varphi_\theta}^{SHARP}(x_i, y_{i1}, y_{i2})$
    If using W-SHARP method, compute $\alpha_{\varphi_\theta}^{W-SHARP}(x_i, y_{i1}, y_{i2})$
  end for
  Let B_L be the top-b elements of B by the value of the acquisition function α.
  Request the preferences labels from the expert and add them to $\mathcal{D}_L$
  Update the policy φ_θ using a gradient step of the $\mathcal{L}_{\text{DPO}}$ using B_L
end for
return $\mathcal{D}_L$ and φ_θ.

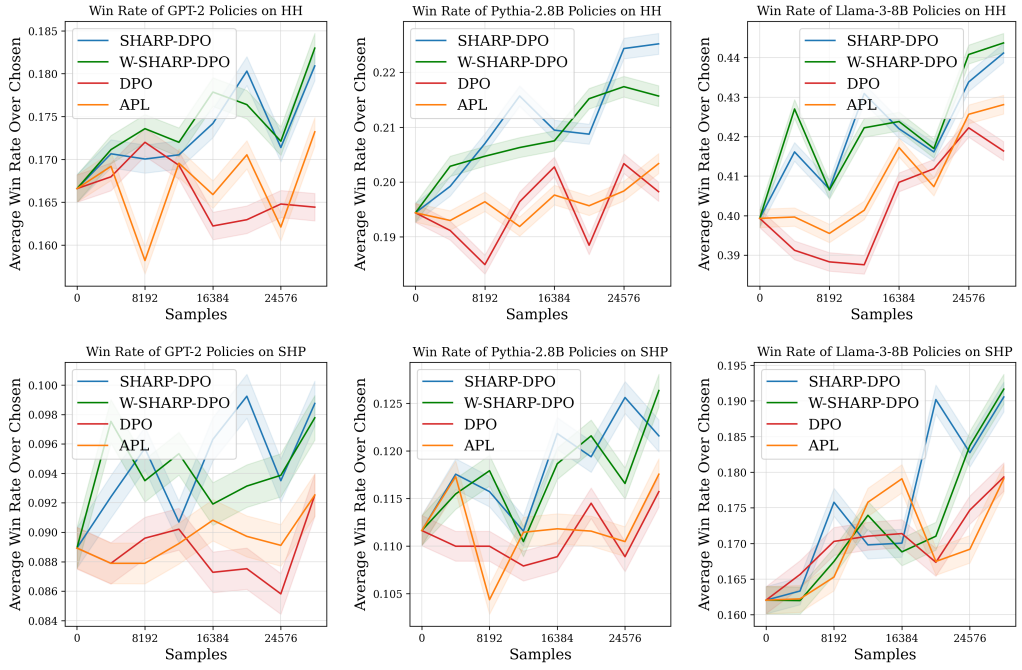

Figure 3: Comparison of W-SHARP-DPO and SHARP-DPO against DPO and APL across different models and datasets. The metric is the average win rate over chosen completions, computed with GPT-4o under swapped evaluation orders to reduce positional bias. Error bars indicate the standard error.

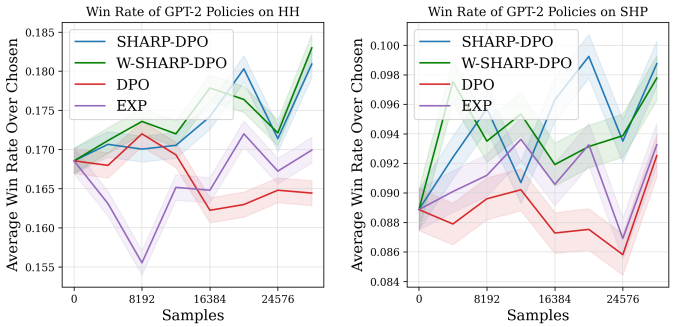

Figure 4: Comparison against the expectation over the gradient on the GPT-2 model with both datasets.

6 Experiments

In this section, we provide the details of our evaluation pipeline. Our main goal is to determine if we can achieve better or comparable performance as DPO while using a smaller amount of labeled data. The standard DPO uses a random selection from the dataset. To assess whether our approaches enhance data selection in DPO, we conduct experiments training large language models (LLMs) on two datasets applied to three different models with different ranges of sizes. We additionally provide comparison results against the APL baseline proposed by Muldrew et al. (2024). When comparing the approaches, we keep all parameters of the experiments identical except for the data selection method to isolate and verify its impact on performance. The code for our approach is publicly available `github.com/belakaria/sharpe-ratio-active-llm-alignment-dpo`.

Datasets We evaluate both methods on two public datasets: the Anthropic Helpful-Harmless (HH) dataset (Bai et al., 2022b) and the Stanford Human Preferences (SHP) dataset (Ethayarajh et al., 2022).

Anthropic Helpful-Harmless (HH): The HH dataset is designed to measure an AI assistant's ability to be both helpful and harmless. It contains two main types of examples: queries where the user's request is reasonable and the assistant should provide a helpful response, and queries where the user's request may be harmful or inappropriate, requiring the assistant to prioritize safety by giving a non-harmful response.

Stanford Human Preferences (SHP): The SHP dataset consists of Reddit posts and corresponding human-generated comments spanning 18 different categories. This broad coverage provides diverse human writing styles and topics. SHP focuses on modeling general human preferences across a wide range of real-world conversations.

LLMs: We explore the impact of active learning by evaluating three models of varying size and capacity: GPT-2, Pythia-2.8-B, and Llama-3-8B. These models span a broad range of resource requirements and capabilities, allowing us to assess how active learning strategies perform under different constraints. We conduct six distinct experiments using the above datasets to provide a comprehensive analysis of each model's performance.

Pipeline Setup: In the DPO pipeline, we begin by splitting each dataset into training and test sets. During the Supervised Fine-Tuning (SFT) phase, we finetune each model on the training split, updating all parameters in each gradient step. In the subsequent DPO phase, to efficiently manage computational resources, we apply a quantized Low-Rank Adaptation (LoRA) of each LLM. This approach reduces memory footprint and speeds up experimentation without sacrificing the model's overall performance. We apply 4-bit quantization using a double quantization strategy under the NF4 scheme while computing in bfloat16. In addition, we use a LoRA configuration with rank 16, alpha 32, and a dropout rate of 0.05, tailored for causal language modeling tasks and omitting additional bias. We set the batch size of our training to 64 and set the fraction that would be labeled to $p = 6$.

We evaluate model performance using the winrate against the dataset's designated "chosen" completions. Formally, the winrate indicates the proportion of generated responses that are

deemed preferable to those labeled as chosen in the dataset. We recompute this metric after every 4,096 training samples to track performance trends over time.

To underscore the benefits of active learning under constrained resources, we limit the DPO phase to a total of 28,672 training points across all experiments. Additionally, we use GPT-4o as an evaluation oracle to compare each newly generated response against the dataset's designated chosen completions. To mitigate position bias, each pair of responses is evaluated twice with reversed ordering, and we report the average winrate across these two evaluations.

Both W-SHARP-DPO and SHARP-DPO consistently outperform the standard DPO baseline and the APL baseline (Figure 3). We attribute this improvement to our acquisition function α, which takes the risk (i.e., all possible gradient outcomes) into account when selecting data points. Interestingly, W-SHARP-DPO and SHARP-DPO achieve similar performance, suggesting that incorporating the implicit reward model as a prior does not necessarily yield further gains in this setting. This could indicate that while using a prior might help in other contexts, it is not required for effective data selection and making no prior assumption could be beneficial for risk assessment.

The accuracy of the implicit reward model for experiments conducted on both datasets echos this result (Figure 5, Appendix). Although this metric is not our primary focus, the results indicate that both SHARP and W-SHARP tend to attain higher accuracy more quickly on the test data, suggesting that these methods guide the model toward more effective reward predictions.

Ablation Study: We additionally provide an ablation study to compare against the expected gradient (Figure 4). We conduct the experiments on the GPT-2 model using both datasets. We observe that the SHARP method performs better. Notably, we expect the performance gap to be even larger in noisy datasets that include response pairs where one option is unlikely or nonsensical. Due to the high memory and computational demands of extracting individual gradients for each data point, we were only able to run these experiments with the GPT-2 model. This computational challenge further underscores the strength of the SHARP approach, as its closed-form expression avoids the need for expensive per-sample gradient computations.

7 Summary, Future Directions, and Limitations

We present a novel active learning strategy for RLHF/DPO in large language models, designed to prioritize and label the most impactful data points under limited human annotation budgets. Central to our method is the use of a Sharpe ratio-based acquisition function to evaluate potential gradient updates. By selecting examples with the highest Sharpe ratios, we aim to target those most likely to produce substantial improvements in policy performance. Our empirical results suggest that this risk-aware selection can reduce annotation costs while enhancing the quality of the learned policy.

Our current approach focuses exclusively on high Sharpe ratio data, which may bias the distribution of selected examples. Although such selective sampling is typical in active learning scenarios, if a practical setting requires an unbiased estimate of the underlying data distribution, future methods could address potential deviations arising from risk-based sampling. Potentially, future methods could combine our Sharpe ratio-based approach with techniques like importance sampling or explicit expectation balancing to address such requirements. Moreover, our computational study was limited by relatively modest resources, restricting the scale of DPO training and the range of datasets evaluated. While our findings demonstrate the promise of a Sharpe ratio-based framework, additional investigation across larger tasks and more extensive experiments would establish its robustness and generalizability.

Although in this work we focused on Sharpe ratio, given its intuitive trade-off between the expected benefit (mean gradient magnitude) and the potential downside (variance), alternative risk-aware metrics such as the Sortino ratio could offer interesting inductive biases. Investigating the impact of such metrics is a promising direction for future work.

References

Azar Alizadeh, Pooya Tavallali, Mohammad Reza Khosravi, and Mukesh Singhal. Survey on recent active learning methods for deep learning. 2021. URL https://api.semanticscholar.org/CorpusID:245763828.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022a. URL https://arxiv.org/abs/2204.05862.

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*, 2022b.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022c.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Rui M. Castro and Robert D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54:2339–2353, 2007. URL https://api.semanticscholar.org/CorpusID:2877584.

Rui M. Castro, Rebecca M. Willett, and Robert D. Nowak. Faster rates in regression via active learning. In *Neural Information Processing Systems*, 2005. URL https://api.semanticscholar.org/CorpusID:1343236.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, A. Kalyan, Karthik Narasimhan, A. Deshpande, and Bruno Castro da Silva. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *ArXiv*, abs/2404.08555, 2024. URL https://api.semanticscholar.org/CorpusID:269137670.

Paul Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinformcement learning from human preferences. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017a. URL https://papers.nips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

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, 2017b.

Paul Doucet, Benjamin Estermann, Till Aczel, and Roger Wattenhofer. Bridging diversity and uncertainty in active learning with self-supervised pre-training. *arXiv preprint arXiv:2403.03728*, 2024.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022.

Teodor Fredriksson, David Issa Mattos, Jan Bosch, and Helena Holmström Olsson. Data labeling: An empirical investigation into industrial challenges and mitigation strategies. In Maurizio Morisio, Marco Torchiano, and Andreas Jedlitschka (eds.), *Product-Focused Software Process Improvement*, pp. 202–216, Cham, 2020. Springer International Publishing. ISBN 978-3-030-64148-1.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Siyi Gu, Minkai Xu, Alexander Powers, Weili Nie, Tomas Geffner, Karsten Kreis, Jure Leskovec, Arash Vahdat, and Stefano Ermon. Aligning target-aware molecule diffusion models with exact energy optimization. *Advances in Neural Information Processing Systems*, 37:44040–44063, 2025.

Hossein Hadian and Hossein Sameti. Active learning in noisy conditions for spoken language understanding. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1081–1090, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://aclanthology.org/C14-1102/.

Steve Hanneke and Liu Yang. Negative results for active learning with convex losses. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 321–325, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/hanneke10a.html.

Kaixuan Ji, Jiafan He, and Quanquan Gu. Reinforcement learning from human feedback with active queries. *arXiv preprint arXiv:2402.09401*, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Forty-first International Conference on Machine Learning*.

Ang Li, Qiugen Xiao, Peng Cao, Jian Tang, Yi Yuan, Zijie Zhao, Xiaoyuan Chen, Liang Zhang, Xiangyang Li, Kaitong Yang, Weidong Guo, Yukang Gan, Xu Yu, Daniell Wang, and Ying Shan. Hrlaif: Improvements in helpfulness and harmlessness in open-domain reinforcement learning from ai feedback, 2024. URL https://arxiv.org/abs/2403.08309.

Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. Sample efficient reinforcement learning from human feedback via active exploration. 2023.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*, 2024.

Fredrik Olsson. A literature survey of active machine learning in the context of natural language processing. 2009.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023a.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=HPuSIXJaa9.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning, 2021. URL https://arxiv.org/abs/2009.00236.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. URL https://api.semanticscholar.org/CorpusID:28695052.

Burr Settles. Active learning literature survey. 2009.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

William F. Sharpe. Mutual fund performance. *The Journal of Business*, 39(1):119–138, 1966. ISSN 00219398, 15375374. URL http://www.jstor.org/stable/2351741.

William F Sharpe. The sharpe ratio. *Streetwise–the Best of the Journal of Portfolio Management*, 3:169–185, 1998.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf, 2024. URL https://arxiv.org/abs/2310.03716.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *ArXiv*, abs/2209.13085, 2022. URL https://api.semanticscholar.org/CorpusID:252545256.

N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize from human feedback. 2020.

Simon Tong and Daphne Koller. Active learning for parameter estimation in bayesian networks. In *Neural Information Processing Systems*, 2000. URL https://api.semanticscholar.org/CorpusID:2386340.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Devis Tuia, Michele Volpi, Loris Copa, Mikhail Kanevski, and Jordi Munoz-Mari. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, 2011. doi: 10.1109/JSTSP.2011.2139193.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Haoran Wang, Qiuye Jin, Shiman Li, Siyu Liu, Manning Wang, and Zhijian Song. A comprehensive survey on deep active learning in medical image analysis. *arXiv preprint arXiv:2310.14230*, 2023. URL https://doi.org/10.48550/arXiv.2310.14230.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

Yuzi Yan, Xingzhou Lou, J. Li, Yipin Zhang, Jian Xie, Chao Yu, Yu Wang, Dong Yan, and Yuan Shen. Reward-robust rlhf in llms. *ArXiv*, abs/2409.15360, 2024. URL https://api.semanticscholar.org/CorpusID:272831831.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *CVPR*, pp. 8941–8951, 2024. URL https://doi.org/10.1109/CVPR52733.2024.00854.

Shenao Zhang, Donghan Yu, Hiteshi Sharma, Ziyi Yang, Shuohang Wang, Hany Hassan, and Zhaoran Wang. Self-exploring language models: Active preference elicitation for online alignment. *arXiv preprint arXiv:2405.19332*, 2024.

Appendix

A Additional Results

In this section, we provide additional results reporting the accuracy of the implicit reward model on the test set. To provide informative results, we use exponential moving average smoothing.

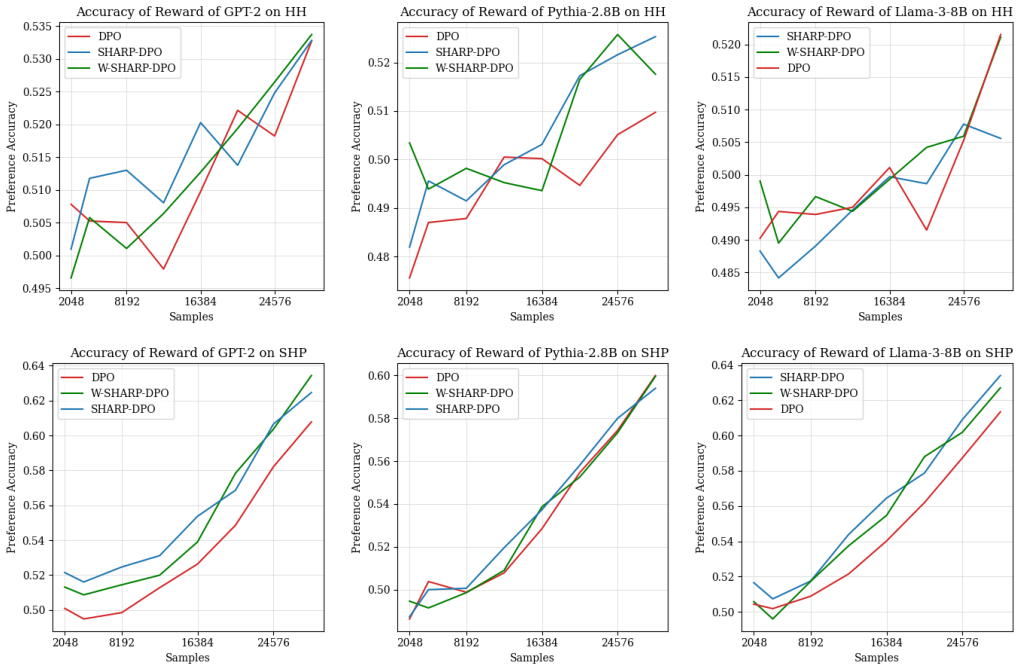

Figure 5: The accuracy of the implicit reward model for models.