Energy-Based Preference Model Offers Better Offline Alignment than the Bradley-Terry Preference Model

Yuzhong Hong [*1] **Hanshan Zhang** [*2] **Junwei Bao** [1] **Hongfei Jiang** [1] **Yang Song** [1]

Abstract

Since the debut of DPO, it has been shown that aligning a target LLM with human preferences via the KL-constrained RLHF loss is mathematically equivalent to a special kind of reward modeling task. Concretely, the task requires: 1) using the target LLM to parameterize the reward model, and 2) tuning the reward model so that it has a 1:1 *linear relationship* with the true reward. However, we identify a significant issue: the DPO loss might have multiple minimizers, of which only one satisfies the required linearity condition. The problem arises from a well-known issue of the underlying Bradley-Terry preference model: it does *not always* have a unique maximum likelihood estimator (MLE). Consequently, **the minimizer of the RLHF loss might be unattainable because it is merely one among many minimizers of the DPO loss**. As a better alternative, we propose an energy-based preference model (EBM) that *always* has a unique MLE, inherently satisfying the linearity requirement. To showcase the practical utility of replacing BTM with our EBM in the context of offline alignment, we adapt a simple yet scalable objective function from the recent literature on fitting EBM and name it as **E**nergy **P**reference **A**lignment (EPA). Empirically, we demonstrate that EPA consistently delivers better performance on open benchmarks compared to DPO, thereby validating the theoretical superiority of our EBM.

[*]Equal contribution [1]Zuoyebang Education Technology (Beijing), Co., Ltd [2]StepFun Technology Co., Ltd.. Correspondence to: Yuzhong Hong <eugene.h.git@gmail.com>, Hanshan Zhang <u5975228@anu.edu.au>, Junwei Bao <baojunwei001@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Table 1. The intrinsic evaluation based on the average Pearson coefficient $\in [-1, 1]$) and the average slope-1 linear regression error $\hat{\epsilon}$ shows that EPA renders a closer approximation to the slope-1 linearity than DPO. This is consistent with the extrinsic evaluation based on the Alpaca Eval 2.0 benchmark.

METHOD	PEARSON (↑) SLOPE-1: NA LINEARITY: ✓	$\hat{\epsilon}$ (↓) SLOPE-1: ✓ LINEARITY: ✓	AE2.0 (%, ↑)
DPO	0.4693	5.78	17.43 / 15.24
EPA (1:1:2)	**0.5808**	5.26	19.20 / 19.26
EPA (1:3:2)	0.5754	**5.01**	**21.31 / 20.13**

1. Introduction

Reinforcement Learning with Human Feedback (RLHF) (Christiano et al., 2017) has been widely used to align a large language model (LLM) with human preference. The canonical RLHF objective (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Perez et al., 2022) is defined as follows (given x):

$$\mathcal{L}_{\text{RLHF}} = \\ - \mathop{\mathbb{E}}_{\pi_\theta(y|x)}[r_{\text{true}}(x,y)] + \beta \text{KL}[\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)] \quad (1)$$

where $\pi_\theta(y|x)$ is the target LLM (i.e., the policy) to tune, $\pi_{\text{ref}}(y|x)$ a frozen LLM initialized identically as the target LLM and $r_{\text{true}}(x,y)$ a reward to maximize.

The $\mathcal{L}_{\text{RLHF}}$ as defined above is not differentiable w.r.t θ (Ziegler et al., 2019; Rafailov et al., 2023), hence not SGD-friendly. Luckily, it has been shown that the unique minimizer of $\mathcal{L}_{\text{RLHF}}$ can be analytically expressed (Korbak et al., 2022a). Then, Rafailov et al. (2023) further reformulate the analytical minimizer as the unique solution to the following set of equations:

$$\begin{cases} r_\theta(x,y) = \beta \log \dfrac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} & (2) \\ r_\theta(x,y) = r_{\text{true}}(x,y) + C(x) & (3) \end{cases}$$

Eq.(2) defines r_θ as the *log ratio reward* and Eq.(3) states that there holds a *slope-1 linearity* between the log ratio reward and the true reward. This formulation implies that as

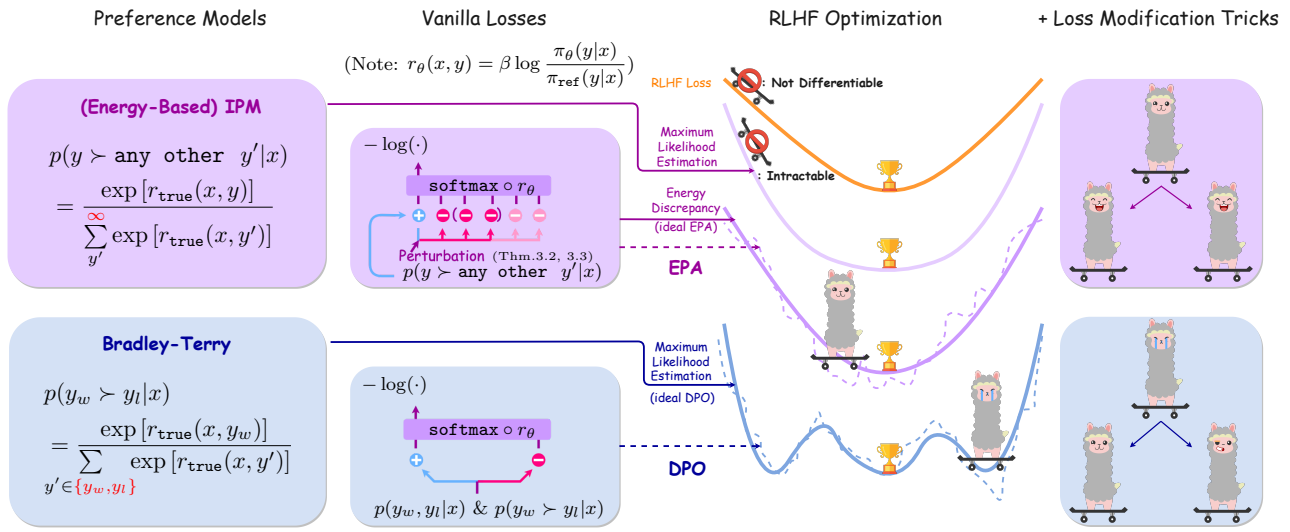

Figure 1. An illustration of the contributions of the paper. Our core argument is that an Energy-Based model (EBM) is a better alternative to the Bradley-Terry model (BTM) due to its guaranteed unique existence of *maximum likelihood estimator* (MLE) (which is identical to the minimizer of the RLHF loss). The advantage of our EBM comes from its intrinsic consideration of the infinity in the size of the space of $y|x$, whereas BTM ignores issues caused by the pair sampling distribution ($p(y_w, y_l|x)$) in such infinite space. Hence we name our EBM the *Infinite Preference Model*. Although approximating the MLE with our proposed EPA loss introduces inevitable error in practice, we find that it is still empirically better performing than its counterpart – DPO, with or without loss modification techniques presented in previous offline alignment literatures.

long as we can find a differentiable objective function $\mathcal{L}(r_\theta)$ to achieve Eq.(3) and r_θ is parameterized by Eq.(2), we can convert the RLHF problem into an offline supervised task. This is the approach of interest in this paper, a drastically different one from classical online RL methods such as PPO.

The poster child example of this offline approach is DPO (Rafailov et al., 2023):

$$\mathcal{L}_{\text{DPO}}^{\text{ideal}}(r_\theta) = -\mathop{\mathbb{E}}_{p(y_w,y_l|x)} \mathop{\mathbb{E}}_{p(y_w \succ y_l|x)} [\log \sigma(r_\theta(x,y_w) - r_\theta(x,y_l))] \tag{4}$$

where y_w and y_l are two responses and $y_w \succ y_l$ means y_w is prefered to y_l given the prompt x. The ideal[1] DPO loss is essentially the maximum likelihood estimation loss of the Bradley-Terry model (BTM) who posits a sigmoidal relationship between $p(y_w \succ y_l|x)$ and $r_{\text{true}}(x,y_w) - r_{\text{true}}(x,y_l)$. *If the maximum likelihood estimator (MLE) uniquely exists*, Rafailov et al. (2023) show that the MLE will make the slope-1 linearity hold.

However, as alluded to by our framing, we argue that **it is *false* to conclude that the slope-1 linearity (i.e., the minimizer of the RLHF loss) is guaranteed to be reached with DPO**. The reason is that the unique existence of BTM's MLE (i.e., the minimizer of $\mathcal{L}_{\text{DPO}}$) is not guaranteed without some non-trivial constraints on the structure of $p(y_w, y_l|x)$, a well-known issue of BTM given an infinite candidate space (i.e., that of $y|x$) in the literature on *learning to rank* (Ford, 1957; Simons & Yao, 1999; Han et al., 2020; Hendrickx et al., 2020; Bong & Rinaldo, 2022; Wu et al., 2022). It is also related to the theoretical issues around *dataset coverage* in the RL literature (Kakade & Langford, 2002; Munos & Szepesvári, 2008; Zhan et al., 2022). Moreover, Tang et al. (2024) have shown that when offline data are drawn from π_{ref} (a usual practice), any pair-wise loss will cease to correlate with $\mathcal{L}_{\text{RLHF}}$ when π_θ deviates enough from π_{ref} due to reward maximization.

These theoretical facts motivate us to find an alternative model that always has a unique MLE that guarantees the slope-1 linearity. As illustrated in Figure 1 and showcased by the results of a proof-of-concept experiment in Table 1, we argue that an Energy-Based model (EBM) called the *Infinite Preference Model* (IPM) is such an alternative. Also, even though its MLE can only be estimated far from perfectly, we observe a closer approximation to the slope-1 linearity and better overall alignment with human preference than DPO. In summary, our contributions are as follows:

- theoretically, we show BTM does not and IPM does

[1] "ideal" in the sense that the expectations in its loss function are accurately computed. In practice, they can only be approximated, which can introduce an additional error.

have guaranteed unique existence of its MLE, equivalent to the minimizer of the RLHF loss;

- empirically, by only loosely estimating the MLE using a contrastive loss called EPA, we achieve a new state of the art of offline alignment on open benchmarks in various settings.

2. Related Work

2.1. DPO and its Recent Improvements

The first approach to avoid DPO's theoretical issue is to use non-BTMs to model data distributions. Rafailov et al. (2023) suggest the DPO's counterpart for the Plackett-Luce Model (we refer to it as DPO-PL), which is a generalized version of BTM for K-wise comparison. IPO (Azar et al., 2023) uses a different pair-wise preference model than BTM. The loss derived from that model can be interpreted as: the difference of log ratio rewards of the y_w and y_l regresses to a constant. However, Tang et al. (2024) show that IPO is still incapable of optimizing $\mathcal{L}_{\text{RLHF}}$, similar to DPO. Ethayarajh et al. (2024) (KTO) point out some limitations of modeling human preference with a pair-wise model. Instead, they independently model a data distribution for desirable samples and another one for undesirable samples. However, such data distributions do not reflect how most benchmark datasets are sampled. This could be the reason why some empirically driven studies find that KTO underperforms DPO on these benchmarks (Meng et al., 2024; Zhou et al., 2024).

The second approach is to tweak the DPO loss. Some loss-tweaking tricks can be effective on their own. For example, Park et al. (2024) (R-DPO) introduce a length penalty on the log ratio reward to make DPO less prone to the verbosity bias. Amini et al. (2024) (ODPO) add a dynamic margin between y_w and y_l based on the intuition that some pairs have stronger or weaker desirability gaps than others. The most effective one discovered so far is on-policy weighting (WPO) (Zhou et al., 2024). Its idea is to approximate the on-policy training scenario by assigning larger weights to the loss of samples closer the current policy at each step and smaller weights to that of less closer ones. Other tricks come in combinations. For example, CPO (Xu et al., 2024) removes the reference model in the log ratio reward and add an SFT loss at the same time. ORPO (Hong et al., 2024) is an improvement over CPO by adding yet another set of tricks: normalizing the policy to the token level (length normalization) and then contrasting the policy distribution with one minus itself. To separate the wheat from the chaff, Meng et al. (2024) find the most simple and effective recipe: removing the reference model, adding a margin and applying length normalization, which gives rise to SimPO. The problem with applying tricks is that there is usually a lack of theoretical justification on how they are related to the minimizer of the RLHF loss.

2.2. Fitting Discrete EBMs

To provide a theoretical background for our proposal, we give a concise review of the most related work on fitting discrete EBMs.

Energy-based models (EBM) (LeCun et al., 2006) are generative models that posit a Boltzmann distribution of data, i.e., $p(x) \propto \exp(-E(x))$ where $E(x)$ (called the *energy function*) is a real-valued function to learn. An EBM is called discrete when the data point x is defined on a discrete space. To fit $p(x)$ with maximum likelihood estimation requires the computation of the normalizer $\Sigma_{x'}^{\infty} \exp(-E(x'))$ (called the *partition function*), which is intractable. Therefore, EBMs are usually learned with a tractable approximation.

The classical approach is to approximate the gradient of maximum likelihood estimation by online sampling from parameterized $p_\theta(x)$ with MCMC (Song & Kingma, 2021). Although they are ideally effective, it is usually difficult or expensive to do such sampling, which harms practical results. Therefore, there are also many MCMC-free methods (Meng et al., 2022; Hyvärinen, 2007; Dai et al., 2020; Lazaro-Gredilla et al., 2021; Eikema et al., 2022). Recently, Schröder et al. (2023) have introduced the notion of energy discrepancy, whose unique global minimizer is identical to the MLE of the EMB in question. Hence, to find the MLE, one can simply minimize the energy discrepancy, which is feasible with SGD on offline data. For its simplicity, we derive EPA based on their theoretical framework.

2.3. EBMs for RLHF

EBMs are not rare in the RLHF research. One of the research directions is to formulate RLHF as a Distribution Matching problem. The typical example for this approach is Distributional Policy Gradients (DPG) (Parshakova et al., 2019; Khalifa et al., 2021). However, we would like to point out that our EBM is different and used for a different purpose. The EBM in DPG is a non-parametric one predefined as the learning signal. Our EBM is a parametric one to fit the distribution of data. The only connection between the two EBMs is that they are used to find the same optimal policy (Korbak et al., 2022a). Deng et al. (2020) uses an EBM for language modeling. Their work essentially solves the self-play-like RLHF problem (Chen et al., 2024b). Although their EBM is also parametric, it fits the optimal policy distribution. Our EBM instead fits the preference distribution. Chen et al. (2024a) proposes two methods – infoNCA and NCA, based on the same EBM as that of Deng et al. (2020). The NCA loss follows the same derivation of the loss proposed by Deng et al. (2020) except that they parameterize

their energy function differently. The infoNCA loss exhibits similarity to our loss. However, we will show that infoNCA is just a worse-performing ablation version of EPA.

3. IPM: Our EBM for Preference Modelling

In the first subsection, we show that an energy-based model (EBM) is guaranteed to have a unique MLE which is equivalent to the minimizer of the RLHF objective. In the second subsection, based on a framework by Schröder et al. (2023), we describe a general strategy to approximate the MLE using offline data. Using it will provide the theoretical account for our proposal in section 4.

3.1. Theoretical Guarantee

Given any x, it is obvious that the space of $y|x$ is infinitely large because y can be any token sequence of unlimited length no matter how likely or unlikely it is a response to x. This infinity is problematic for BTM. For example, if there is a single y that is never sampled, it is easy to refute the unique existence of BTM's MLE (see Proposition B.5). However, EBM can naturally take the infinity into account to avoid the issue. Specifically, we model a one-to-infinite preference (v.s. BTM and the more general Plackett-Luce model only model a one-to-finite-number preference) as follows:

$$p(y|x) \coloneqq p(\forall y' \neq y : \ y \succ y'|x) \tag{5}$$

Namely, $p(y|x)$[2] is the probability that candidate y is preferred over all other candidates. Under mild assumptions (Assumptions B.1 and B.2) that make an EBM applicable, we define the Infinite Preference Model (IPM) to be the one that posits that $p(y|x)$ is a Boltzmann distribution induced by the corresponding true reward (i.e., using $-r_{\text{true}}(x, y)$ as the energy function):

$$p(y|x) = \frac{\exp[r_{\text{true}}(x, y)]}{\Sigma_{y'}^{\infty} \exp[r_{\text{true}}(x, y')]} \tag{6}$$

IPM is a better alternative to BTM because of the following theorem (see Appendix B for proof).

Theorem 3.1. *when we parameterize the IPM as follows, the unique existence of the IPM's MLE is guaranteed and it will be reached if and only if the slope-1 linearity (i.e., Eq.(3)) holds between the log ratio reward and the true reward.*

$$q_\theta(y|x) = \frac{\exp[r_\theta(x, y)]}{\Sigma_{y'}^{\infty} \exp[r_\theta(x, y')]} \tag{7}$$

where r_θ is defined as in Eq.(2).

[2]One should not confuse $p(y|x)$ with $\pi(y|x)$ although both of them are distributions over y given x. $p(y|x)$ is how likely humans would rate a y as the best whereas $\pi(y|x)$ measures how likely y is to be generated.

Therefore, as long as we can find the MLE of the IPM parameterized as so, we are guaranteed to reach the minimizer of $\mathcal{L}_{\text{RLHF}}$ since it is the unique solution to Eq.(2) and Eq.(3).

On a side note, the IPM has been previously introduced by other studies on RLHF for a different purpose: to theoretically equate the maximization of $-\mathcal{L}_{\text{RLHF}}$ to the variational inference of the optimal policy with π_{ref} as the prior (Korbak et al., 2022b; Yang et al., 2024). However, to the best of our knowledge, we are the first one to introduce IPM not just as a theoretical toy, but as a tool (when parameterized by the log ratio reward) to do proper RL-free RLHF.

3.2. Offline Approximation of MLE

Despite the powerfulness of IPM, finding its MLE is a non-trivial task. Directly finding it with the minimization of the negative log likelihood $-\log q_\theta(y|x)$ is intractable because of the infinity in the denominator.

There are good tractable approximations but usually with complex online training algorithms. For simplicity and scalability purposes, we choose to follow Schröder et al. (2023), who provide a general strategy that finds the optimal EBM by simple SGD with offline training data. The strategy is based on two theorems formally adapted for our purpose as follows.

Theorem 3.2. *For any random variable Z with the conditional variance $\text{Var}[Y|Z]$ being positive, the global unique minimizer $r^*(x, y)$ of the functional* Energy Discrepancy *(ED) defined as follows is the optimal IPM (i.e., $p(y|x) \propto \exp[r^*(x, y)]$).*

$$ED_{x,p(y|x),p(z|y)}[r] \coloneqq \mathop{\mathbb{E}}_{p(y|x)} \mathop{\mathbb{E}}_{p(z|y)} [\log \Sigma_{y'} p(z|y') \exp[r(x, y') - r(x, y)]] \tag{8}$$

Theorem 3.3. *For any random variable Z whose backward and forward transition probabilities from Y solve the equation $\Sigma_y p(z|y) f(y) = \Sigma_y p(y|z) f(y)$ for an arbitrary f, the estimation error of the following statistic estimate of $ED_{x,p(y|x),p(z|y)}[r]$ vanishes almost surely when $N \to \infty$ and $M \to \infty$.*

$$\mathcal{L}[\theta|x] \coloneqq \frac{1}{N} \Sigma_i^N \log(1 + \Sigma_j^M \exp[r_\theta(x, y_-^{i,j}) - r_\theta(x, y^i)]) - \log(M) \tag{9}$$

where $\{y^i\}^N$ are samples from $p(y|x)$, $\{y_-^{i,j}\}^M$ from $p(y|z_0)$, and z_0 a single sample from $p(z|y^i)$.

A one-sentence interpretation of the above theorems is: if we have a particular kind of negative sampling strategy by perturbing observed preferred samples, we will learn the optimal IPM by minimizing a contrastive loss between the

negatives and observed positives. Therefore, when the IPM is parameterized by the log ratio reward, we will find the exact minimizer of $\mathcal{L}_{\text{RLHF}}$ with this loss function.

Note that the property of the negative sampling source Z as described in Theorem 3.3 is just a sufficient condition as opposed to a necessary one. This leaves room for empirical discovery of better negative sampling strategies. A rule of thumb as suggested by Schröder et al. (2023) is that Z **has to be informative of** Y **and of high conditional variance at the same time**. This provides the intuition of our proposal in section 4.

4. EPA: A Practical Approximation

To introduce our loss, we first write the ideal loss in Eq.(9) in an equivalent form by removing the constant $\log(M)$ and moving a minus sign out of the logarithm:

$$\tilde{\mathcal{L}}[\theta|x] \coloneqq \frac{1}{N}\Sigma_i^N - \log \frac{\exp[r_\theta(x, y^i)]}{\exp[r_\theta(x, y^i)] + \Sigma_j^M \exp[r_\theta(x, y_-^{i,j})]} \tag{10}$$

Now we propose our loss function in this *negative log softmax* form with a specific negative sampling strategy in mind.

4.1. Narrow Definition

For the most classical setup, we assume we only have access to pair-wise preference data. In this setting, our loss for each mini-batch of B samples ($\{(x^i, y_w^i, y_l^i)\}^B$) is defined as follows:

$$\mathcal{L}_{\text{EPA}} = \frac{1}{B}\sum_i^B - \log \frac{\exp[r_\theta(x^i, y_w^i)]}{\sum_{j\in\{i\}\cup\mathcal{I}_{\text{wk}}} \big(\exp[r_\theta(x^i, y_w^j)] + \exp[r_\theta(x^i, y_l^j)]\big)} \tag{11}$$

where $\mathcal{I}_{\text{wk}}$ is a non-empty random subset of $\{1,\ 2,\ \ldots,\ i-1,\ i+1,\ \ldots,\ B\}$, introducing negative samples that are mismatched responses originally sampled for other prompts. Its size $|\mathcal{I}_{\text{wk}}| = N_{weak}^-/2 \in (0, B-1]$ is a hyperparameter. Note that our loss without $\mathcal{I}_{\text{wk}}$ reduces to the DPO loss. We justify our choice of positives and negatives in EPA as follows:

1. **Why is y_w a good approximation of a positive sample from $p(y|x)$?** For a y_w in the dataset, it may not be the best y, but there is only a finite number of potentially possible better ones according to Assumption B.1. Also, since we know it is preferred over y_l and infinitely many other arbitrary token sequences, it is a good approximation of a y that is preferred over all other samples up to a small error.
2. **Why use both y_l and mismatched responses as negatives?** As stated at the end of section 3, we want to draw the negatives from a perturbation source that is both informative of the positives and of high variance at the same time. For the informativeness, we consider *strong* negatives y_l because they are semantically close to y_w. For the high variance, we consider *weak* negatives such as mismatched responses. We will show the effectiveness of such choice with ablation experiments in section 5.

4.2. General Definition

Note that the number of strong negatives in Eq.(11) is limited to 1 because of the given pair-wise data. This is not ideal for the approximation of IPM's MLE because the number of negatives should be large enough to reduce the approximation error (Theorem 3.3). Therefore, we generalize our definition of EPA to circumstances where we can have access to more strong negatives (i.e., each y_w is accompanied by $\{y_{l_1},\ y_{l_2},\ \ldots\}$ instead of just one y_l). This is practically feasible because we can always sample less desirable responses from some LLM.

Hence, we define our loss in a more general form as follows:

$$\mathcal{L}_{\text{EPA}}^{\text{general}} = \frac{1}{B}\sum_i^B - \log \frac{\exp[r_\theta(x^i, y_w^i)]}{\exp[r_\theta(x^i, y_w^i)] + \sum_{k\in\mathcal{I}_{\text{st}}} \exp[r_\theta(x^i, y_{l_k}^i)] + \sum_{j\in\mathcal{I}_{\text{wk}}} \exp[r_\theta(x^i, y_*^j)]} \tag{12}$$

where $\mathcal{I}_{\text{st}}$ contains N_{strong}^- indices of available strong negatives; $\mathcal{I}_{\text{wk}}$ contains N_{weak}^- indices of weak negatives; y_*^j can be either y_w^j or some $y_{l_k}^j$ ($j \neq i$).

4.3. Gradient Analysis

Using the chain rule, one can easily find the gradient of the EPA loss (the general one) as follows:

$$\nabla_\theta \mathcal{L}_{\text{EPA}}^{\text{general}} = -\frac{\beta}{B}\sum_i^B \Big(\underbrace{\sum_{k\in\mathcal{I}_{\text{st}}} s_{l_k}^i \big(\nabla_\theta \log \pi_\theta(y_w^i|x^i) - \nabla_\theta \log \pi_\theta(y_{l_k}^i|x^i)\big)}_{\text{strong contrast}} + \underbrace{\sum_{j\in\mathcal{I}_{\text{wk}}} s^j \big(\nabla_\theta \log \pi_\theta(y_w^i|x^i) - \nabla_\theta \log \pi_\theta(y_*^j|x^i)\big)}_{\text{weak contrast}} \Big) \tag{13}$$

where $s_{l_k}^i$ and s^j are the softmax-ed values of the strong negative log ratio rewards and the weak negative log ratio rewards, respectively. They control the magnitude of the strong and weak contrast. When there is no weak contrast, the gradient reduces to that of the DPO loss if there is only one strong negative. Therefore, one can interpret the weak

contrast as a regularization term added to DPO to prevent θ from moving to a direction that undesirably increases the likelihood of weak negatives.

5. Experiments

5.1. Experimental Setup

5.1.1. Training Data

We consider the dataset of Ultrafeedback (Cui et al., 2024) (denoted as 'UF-all') and a widely used pair-wise version of it (Tunstall et al., 2023) (UF-binarized). The two datasets are ideal for our purpose besides their popularity. Firstly, in UF-all, there are 4 responses sampled from multiple sources for each prompt. This will allow training with our general version of EPA which can utilize multiple strong negatives. Secondly, in both UF-all and UF-binarized, the positive sample y_w for each x is the best one out of the 4 responses. This is an arguably close approximation to our assumption that positives are sampled from $p(\forall y' \neq y : \ y \succ y'|x)$.

5.1.2. Evaluation

Although Ultrafeedback is intended to reflect human preference, it is labeled by GPT-4 in reality. Consequently, we consider MT-Bench (Zheng et al., 2024) which also uses GPT-4 to score a response on a scale of 1-10. The metric is the average score for 80 single-turn conversations and 80 multi-turn conversations. We also consider Alpaca-Eval 2.0 (Dubois et al., 2024) because of its high correlation with human preference, the ultimate concern for RLHF. Its metrics are win-rates (with or without *length control*) against GPT-4-turbo across 805 test samples with the judge being GPT-4-turbo itself. We report them in the format of "length controlled win-rate / win-rate" in the experiment results. For evaluation on Alpaca Eval 2.0, we use the default decoding parameters in the Huggingface implementation. For MT-Bench, we use the ones specifically required by the benchmark.

5.1.3. Baselines & Loss Modification Tricks

For fair comparison, we only consider methods from the approach that explicitly aims to minimize $\mathcal{L}_{\text{RLHF}}$ with specific probabilistic models about data distributions. Therefore, we consider DPO, IPO, KTO and NCA for the classical pair-wise data setting. We consider DPO-PL, NCA, and infoNCA for the general setting where there are multiple responses for each prompt in the dataset.

Loss modification tricks are not considered as baselines because they are *orthogonal* to our proposal. Comparing BTM+tricks to EBM would be comparing apples to oranges. Instead, we consider applying the tricks to both EPA (the narrow one for fair comparison) and DPO to further verify our core argument about EBM's superiority over BTM. The tricks in consideration are those used in SimPO, R-DPO, CPO and WPO (Details in Table 7 in the Appendix C).

5.1.4. Implementation

We use mistral-7b-sft-beta[3] as the reference model and for the initialization of policy in our paper. We train all models in this paper for 3 epochs with LoRA ($r = 16$, $\alpha = 16$, dropout $= 0.05$). Whenever comparing among different methods, we pick the one out of the three checkpoints with the best MT-Bench score for each method. For fair comparison of baseline models, we fix β to 0.01. It is more of a *control variable* than a hyperparameter because it is a given component of the RLHF objective which all baselines are aimed to optimize. We only vary β for them when probing their KL-Reward frontiers. For comparison of loss modification tricks, since the RLHF objective is not necessarily the purpose, we use the best β and other hyperparameters specific to each method as reported in previous work (e.g., the tricks used in SimPO are only competitive when $\beta = 2.0$ for the Mistral model). Learning rate is grid-searched for each method among $\{1e-5, 5e-6, 1e-6\}$.

Table 2. EPA beats all other baselines either for pair-wise data or for data with > 2 responses for each prompt.

Training Data	Method	AE 2.0 (%)	MT-Bench
	SFT	8.16 / 5.47	6.44
UF-binarized	+DPO	17.43 / 15.24	7.55
	+IPO	12.97 / 10.13	7.31
	+KTO	12.62 / 11.29	7.21
	+NCA	14.64 / 11.27	7.39
	+EPA	**19.20 / 19.26**	**7.71**
UF-all	+DPO-PL	15.95 / 14.68	7.57
	+NCA	15.08 / 11.85	7.28
	+infoNCA	17.30 / 16.25	7.50
	+EPA	**22.03 / 21.44**	**7.58**

5.2. Results and Analysis

5.2.1. EPA Performs Better than Baselines

As shown in Table 2, we can see EPA consistently achieves the highest scores and hence a new state of the art. Note that other baselines generally perform even less well than DPO. This makes BTM the strongest baseline for IPM.

To compare IPM with its most competitive baseline BTM in detail, we come back to the starting point – optimizing $\mathcal{L}_{\text{RLHF}}$. We study from two perspectives of the optimization problem. Both perspectives involve multiple checkpoints

[3] https://huggingface.co/HuggingFaceH4/mistral-7b-sft-beta

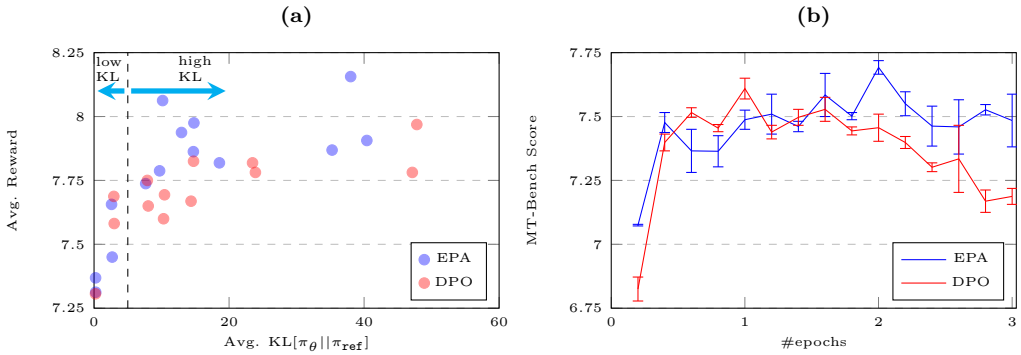

Figure 2. DPO vs. EPA (1:1:2) from the perspective of (a) KL-Reward frontier and (b) training dynamics.

beyond the single best one for each method (e.g., Table 2), offering a more comprehensive comparison.

First, we study how well each method balances the KL term and the reward term in $\mathcal{L}_{\text{RLHF}}$ with varying $\beta \in \{0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.5\}$. Both terms are computed on the 80 single-turn prompts in the MT-Bench dataset. We estimate KL with 20 response samples per prompt from each policy distribution. We use the GPT-4 score produced by MT-Bench as an alias for the true reward. As shown in Figure 2.(a), in the high-KL region, EPA generally achieves higher reward than DPO. The two become indistinguishable only in the low-KL region.

Second, to understand how EPA differs from DPO in terms of the dynamics during the optimization process, we test the MT-Bench score of the checkpoint at every 20% of an epoch. As shown in Figure 2.(b), EPA is less prone to overfitting and fits to the reward signal more steadily than DPO. The performance of DPO starts to degenerate rapidly after the first epoch. However, EPA reaches its peak performance at the end of the second epoch and overfits more slowly afterward. This is consistent with our gradient analysis in Section 4 that EPA is more regularized than DPO.

5.2.2. Combining Strong and Weak Negatives is Effective

We also run ablation and variants of EPA for different numbers of strong and weak negatives (ie., two hyperparameters of EPA). As shown in Table 5.2.2, we can observe the pattern that strong negatives and weak negatives should be roughly balanced to achieve the best overall performance. The absence or excessiveness of either kind of the negatives will lead to poorer performance. This is consistent with the intuition that the negative distribution should be both informative and of high variation.

Table 3. Ablation and variants of EPA with varying number of strong negatives (N^-_{strong}) and that of weak negatives (N^-_{weak}) in addition to the 1 positive ($N^+ = 1$) in the denominator of $\mathcal{L}^{\text{general}}_{\text{EPA}}$.

Method	N^+:N^-_{strong}:N^-_{weak}	AE 2.0 (%)	MT-Bench
Ablation	1:1:0 (DPO)	17.43 / 15.24	7.55
	1:0:2	9.37 / 6.74	6.57
EPA	1:1:1	**21.14 / 20.55**	7.29
	1:1:2	19.20 / 19.26	**7.71**
	1:1:6	16.63 / 15.78	7.57
Ablation	1:3:0 (InfoNCA)	17.30 / 16.25	7.50
EPA	1:3:2	22.03 / 21.44	**7.58**
	1:3:4	21.31 / 20.13	**7.58**
	1:3:6	24.01 / 23.44	7.35
	1:3:8	**24.54 / 23.75**	7.19
	1:3:10	23.58 / 22.78	7.43

Table 4. Adding the same number of weak negatives to the data for DPO (as additional y_l to be paired with the original y_w) or DPO-PL (as additional negatives ranked after the y_w and y_l) does not show any advantage over EPA. +UF-weak×1 means to add a set of random weak negatives that is of the same quantity of prompts in UF-binarized. +UF-weak×2 means to add 2 such sets.

Training Data	Method	AE 2.0 (%)	MT-Bench
UF-binarized	DPO	17.43 / 15.24	7.55
	EPA (1:1:1)	**21.14 / 20.55**	7.29
	EPA (1:1:2)	19.20 / 19.26	**7.71**
+ UF-weak×1	DPO	18.99 / 17.00	7.37
	DPO-PL	15.78 / 15.17	7.54
+ UF-weak×2	DPO	18.49 / 18.72	7.42
	DPO-PL	15.59 / 14.79	7.44

Table 5. BTM vs. IPM with empirically effective loss modification tricks.

PREF MODEL	LOSS MODIFICATION	AE 2.0 (%)	MT-BENCH
BTM	N/A (DPO)	17.43 / 15.24	7.55
	$-ref + \mathcal{L}_{sft}$ (CPO)	13.63 / 10.34	7.11
	$+len_p$ (R-DPO)	19.10 / 16.71	<u>7.70</u>
	$-ref + len_n + m_c$ (SIMPO)	20.57 / 20.19	7.61
	$+w_{op}$ (WPO)	21.90 / <u>21.04</u>	7.56
	$+m_c$	<u>22.33</u> / 20.61	7.67
IPM	N/A (EPA)	19.20 / 19.26	**7.71**
	$+w_{op}$	22.80 / 22.26	7.61
	$+m_c$	**23.00 / 22.47**	7.68

5.2.3. COMPUTATION COST-EFFECTIVENESS COMPARED TO ALTERNATIVES

There might be a concern that EPA's effectiveness merely comes from more computations induced by the additional weak negatives. Therefore, we also experiment with the introduction of weak negatives to BTM and the more general Placket-Luce Model. As shown in Table 4, we can see that EPA still performs better than DPO and DPO-PL in the setting where the total number of computations is strictly controlled. This indicates that the effectiveness is not based on the additional computations alone, but essentially a consequence of the superior theoretical property of IPM, an EBM that has a unique MLE. Nevertheless, one can still argue that it is a somewhat shortcoming that EPA has a computation complexity linearly related to the number of contrasting samples. However, when there is an additional computation budget available in practice, EPA is indeed so far the most cost-effective method to fully exploit the resource. In addition, we want to emphasize that the computation cost is not an intrinsic attribute of replacing BTM with IPM (the core contribution/purpose of this paper), but only an attribute of EPA (a not-so-perfect way to estimate the IPM's MLE). As the studies of EBM continue to progress, it is reasonable to expect IPM to outrun BTM even more.

5.2.4. IPM+TRICKS >= BTM+TRICKS

Since most loss modification tricks presented in the recent offline alignment literature are originally intended for BTM/DPO and do not necessarily make sense to IPM/EPA, we only consider two of them when applying to EPA. The first one is a constant margin m_c added to the logit of y_l. The trick can be viewed as a loose numerical approximation to the general EPA where there are multiple y_{l_k}. For example, if $m_c = 1.4$, we have $\exp[r_\theta(x, y_l) + m_c] = \exp[m_c] \times \exp[r_\theta(x, y_l)] \approx 4 \times \exp[r_\theta(x, y_l)]$. The second one is the on-policy weight proposed by Zhou et al. (2024). It can be viewed as a curriculum learning technique which prioritizes samples that closely relate to the current policy distribution at each step.

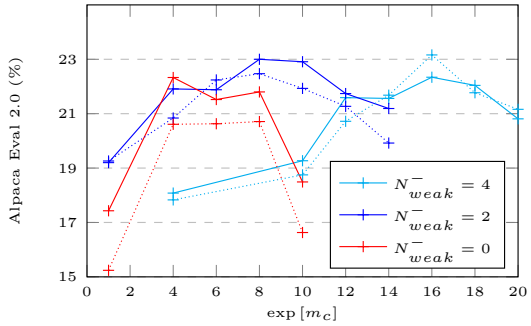

Figure 3. Performance of modified DPO ($N^-_{weak} = 0$) and modified EPA ($N^-_{weak} > 0$) with a margin m_c added to $r_\theta(y_l|x)$. Solid lines represent the length-controlled win-rates, and dotted lines represent the raw win-rates.

As shown in Table 5, we find that both $+w_{op}$ and $+m_c$ produce similar performance boost on EPA to DPO. Although the marginal boost on EPA is generally smaller than DPO, EPA with tricks is still better than DPO with tricks. However, the fact these tricks can still work on EPA also implies that there is still room for improvement. This may come from EPA not necessarily being the best algorithm to approximate IPM's MLE.

We also study in detail how the value of m_c influences DPO and EPA. As shown in Figure 3, we observe that a combination of higher m_c and higher N^-_{weak} tends to produce higher performance. A possible explanation for this is that as m_c scales up the logit of the strong negative y_l to loosely approximate the existence of multiple strong negatives, we get closer to the performance of the general EPA.

6. Conclusion

In this paper, we show both BTM and our EBM (IPM) have the property that their MLE, if uniquely exists, is equivalent to the minimizer of the RLHF loss. However, the unique existence of IPM's MLE is guaranteed whereas that of BTM's MLE is not. This theoretical advantage implies that as long as the IPM's MLE is accurately found, we are bound to minimize the RLHF loss. But, the same claim does not hold for BTM. Although EPA is just an empirical attempt to approximate IPM's MLE, it is already sufficient to outperform its counterpart – DPO on open benchmarks, with or without loss modification tricks presented in previous work.

However, EPA is far from perfect. For example, relatively poorer computation and memory efficiency is a major handi-

cap of EPA. Foreseeable future work includes finding better ways to perturb data or adopting more efficient methods to approximate the MLE. Loss modification tricks particularly tailored for EPA also remain to be explored.

Acknowledgements

We thank Yulong Zhou and Shaoke Lv for many insightful discussions. We also thank our colleagues who provided stable hardware environments for conducting our experiments. Last but not least, we are grateful to the anonymous reviewers for their meaningful suggestions that helped refine this paper.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Amini, A., Vieira, T., and Cotterell, R. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

Bong, H. and Rinaldo, A. Generalized results for the existence and consistency of the MLE in the bradley-terry-luce model. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2160–2177. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bong22a.html.

Chen, H., He, G., Su, H., and Zhu, J. Noise contrastive alignment of language models with explicit rewards. *arXiv preprint arXiv:2402.05369*, 2024a.

Chen, Z., Deng, Y., Yuan, H., Ji, K., and Gu, Q. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024b.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Cui, G., Yuan, L., Ding, N., Yao, G., He, B., Zhu, W., Ni, Y., Xie, G., Xie, R., Lin, Y., et al. Ultrafeedback: Boosting language models with scaled ai feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Dai, H., Singh, R., Dai, B., Sutton, C., and Schuurmans, D. Learning discrete energy-based models via auxiliary-variable local exploration. *Advances in Neural Information Processing Systems*, 33:10443–10455, 2020.

Deng, Y., Bakhtin, A., Ott, M., Szlam, A., and Ranzato, M. Residual energy-based models for text generation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1l4SgHKDH.

Dubois, Y., Galambosi, B., Liang, P., and Hashimoto, T. B. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Eikema, B., Kruszewski, G., Dance, C. R., Elsahar, H., and Dymetman, M. An approximate sampler for energy-based models with divergence diagnostics. *Transactions on Machine Learning Research*, 2022.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization, 2024.

Ford, L. R. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64 (8):28–33, 1957. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2308513.

Han, R., Ye, R., Tan, C., and Chen, K. Asymptotic theory of sparse Bradley–Terry model. *The Annals of Applied Probability*, 30(5):2491 – 2515, 2020. doi: 10.1214/20-AAP1564. URL https://doi.org/10.1214/20-AAP1564.

Hendrickx, J., Olshevsky, A., and Saligrama, V. Minimax rate for learning from pairwise comparisons in the BTL model. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4193–4202. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hendrickx20a.html.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300, 2020. URL https://arxiv.org/abs/2009.03300.

Hong, J., Lee, N., and Thorne, J. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2(4):5, 2024.

Hyvärinen, A. Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512, 2007.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Khalifa, M., Elsahar, H., and Dymetman, M. A distributional approach to controlled text generation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jWkw45-9AbL.

Korbak, T., Elsahar, H., Kruszewski, G., and Dymetman, M. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. *Advances in Neural Information Processing Systems*, 35:16203–16220, 2022a.

Korbak, T., Perez, E., and Buckley, C. Rl with kl penalties is better viewed as bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1083–1091, 2022b.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lazaro-Gredilla, M., Dedieu, A., and George, D. Perturb-and-max-product: Sampling and learning in discrete energy-based models. *Advances in Neural Information Processing Systems*, 34:928–940, 2021.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Li, T., Chiang, W.-L., Frick, E., Dunlap, L., Wu, T., Zhu, B., Gonzalez, J. E., and Stoica, I. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

Majerek, D., Nowak, W., and Zieba, W. Conditional strong law of large number. *Int. J. Pure Appl. Math*, 20(2): 143–156, 2005.

Meng, C., Choi, K., Song, J., and Ermon, S. Concrete score matching: Generalized score matching for discrete data. *Advances in Neural Information Processing Systems*, 35: 34532–34545, 2022.

Meng, Y., Xia, M., and Chen, D. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Park, R., Rafailov, R., Ermon, S., and Finn, C. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

Parshakova, T., Andreoli, J.-M., and Dymetman, M. Distributional reinforcement learning for energy-based sequential models. *Optimization Foundations for Reinforcement Learning Workshop at NeurIPS 2019*, 2019. URL https://optrl2019.github.io/assets/accepted_papers/34.pdf.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, 2022.

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

Schröder, T., Ou, Z., Li, Y., and Duncan, A. B. Training discrete EBMs with energy discrepancy. In *ICML 2023 Workshop: Sampling and Optimization in Discrete Space*, 2023. URL https://openreview.net/forum?id=kFMpJh75Wo.

Simons, G. and Yao, Y.-C. Asymptotics when the number of parameters tends to infinity in the bradley-terry model for paired comparisons. *The Annals of Statistics*, 27(3): 1041–1060, 1999.

Song, Y. and Kingma, D. P. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Tang, Y., Guo, Z. D., Zheng, Z., Calandriello, D., Munos, R., Rowland, M., Richemond, P. H., Valko, M., Pires, B. Á., and Piot, B. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*, 2024.

Tikhonov, A. and Ryabinin, M. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. *CoRR*, abs/2106.12066, 2021. URL https://arxiv.org/abs/2106.12066.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Wu, W., Junker, B. W., and Niezink, N. Asymptotic comparison of identifying constraints for bradley-terry models. *arXiv preprint arXiv:2205.04341*, 2022.

Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Van Durme, B., Murray, K., and Kim, Y. J. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024.

Yang, J. Q., Salamatian, S., Sun, Z., Suresh, A. T., and Beirami, A. Asymptotics of language model alignment, 2024.

Yuan, H., Yuan, Z., Tan, C., Wang, W., Huang, S., and Huang, F. RRHF: Rank responses to align language models with human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=EdIGMCHk4l.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhou, W., Agrawal, R., Zhang, S., Indurthi, S. R., Zhao, S., Song, K., Xu, S., and Zhu, C. Wpo: Enhancing rlhf with weighted preference optimization. *arXiv preprint arXiv:2406.11827*, 2024.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Álvaro Bartolomé Del Canto, Blázquez, G. M., Lajarín, A. P., and Suero, D. V. Distilabel: An ai feedback (aif) framework for building datasets with and for llms. https://github.com/argilla-io/distilabel, 2024.

A. Proof of the Equivalence Between Slope-1 Linearity and Minimizer of the RLHF Loss

We do not claim any originality for the proofs given in this section because they are largely just paraphrased versions of the work by Korbak et al. (2022a); Rafailov et al. (2023) and others. We include them just for reference and completeness of the mathematical foundation shared by both DPO and EPA. Throughout this paper, we make the same mild assumption as by Rafailov et al. (2023) that π_{ref} is strictly positive.

Lemma A.1. *The minimizer of the RLHF objective uniquely exists.*

Proof. We show the minimizer π_r can be analytically expressed by $\frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\frac{1}{\beta}r(x,y)$ where $Z(x) = \Sigma_y^\infty \pi_{\text{ref}}(y|x)\exp\frac{1}{\beta}r(x,y)$ (i.e., a normalizer to make π_r a probabilistic distribution).

From the property of Gibb's inequality, we know:

$$\pi_r = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp[\frac{1}{\beta}r(x,y)]$$
$$\Leftrightarrow$$
$$\pi_r = \arg\min_{\pi_\theta} \beta\text{KL}[\pi_\theta(y|x)||\frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\frac{1}{\beta}r(x,y)]$$

We will complete the proof by showing the βKL-Divergence on the RHS of the above equation is the RLHF objective itself plus a constant w.r.t θ:

$$\begin{aligned}
&\beta\text{KL}[\pi_\theta(y|x)||\frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\frac{1}{\beta}r(x,y)] \\
&= \beta \mathop{\mathbb{E}}_{\pi_\theta(y|x)}[\log\frac{\pi_\theta(y|x)}{\frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\frac{1}{\beta}r(x,y)}] \\
&= \beta \mathop{\mathbb{E}}_{\pi_\theta(y|x)}[\log\frac{Z(x)}{\exp[\frac{1}{\beta}r(x,y)]} + \log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}] \\
&= \beta \mathop{\mathbb{E}}_{\pi_\theta(y|x)}[\log Z(x) - \frac{1}{\beta}r(x,y) + \log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}] \\
&= -\mathop{\mathbb{E}}_{\pi_\theta(y|x)}[r(x,y)] + \beta\text{KL}[\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)] \\
&\quad + \beta\log Z(x) \\
&= \mathcal{L}_{RLHF}(\theta) + \beta\log Z(x)
\end{aligned}$$

□

Definition A.2. We say a slope-1 linearity holds when:

$$r_\theta(x,y) = r(x,y) + C(x)$$

where $r_\theta(x,y) = \beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$.

Theorem A.3 (Theorem of necessity)**.** *If $\pi_\theta = \pi_r$, then slope-1 linearity holds.*

Proof. If $\pi_\theta = \pi_r$, then according to Lemma A.1, we have:

$$\pi_\theta = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\frac{1}{\beta}r(x,y)$$

Take the logarithm of both sides of this equation, we have:

$$\log\pi_\theta = \log\pi_{\text{ref}} + \frac{1}{\beta}r(x,y) - \log Z(x)$$

After moving the two log terms to the same side, we get slope-1 linearity:

$$\beta\log\frac{\pi_\theta}{\pi_{\text{ref}}} = r(x,y) - \beta\log Z(x)$$

□

Theorem A.4 (Theorem of sufficiency)**.** *If slope-1 linearity holds, then $\pi_\theta = \pi_r$.*

Proof. From the Theorem of necessity, we know:

$$\beta\log\frac{\pi_r}{\pi_{\text{ref}}} = r(x,y) - \beta\log Z(x)$$

Substracting the slope-1 linearity from this equation, we get:

$$\beta\log\frac{\pi_r}{\pi_{\text{ref}}} - \beta\log\frac{\pi_\theta}{\pi_{\text{ref}}} = -\beta\log Z(x) - C(x)$$

Eliminating the non-zero β from both sides and taking the exponential, we have:

$$\frac{\pi_r}{\pi_\theta} = f(x)$$

where $f(x) = \frac{1}{Z(x)\exp[\frac{1}{\beta}C(x)]}$. Moving π_θ to the RHS, we get:

$$\pi_r = \pi_\theta f(x)$$

Taking Σ_y^∞ for both sides, we can sum up both distributions π_r and π_θ to one:

$$1 = f(x)$$

Therefore,

$$\pi_r = \pi_\theta f(x) = \pi_\theta$$

□

Note that from the above proof, we can easily get the following corollary because $f(x) = 1 \Leftrightarrow C(x) = -\beta\log Z(x)$.

Corollary A.5. *when a r_θ satisfies slope-1 linearity, it is unique.*

B. Theoretical Aspect of the Infinite Preference Model

We will first give our proof of the guaranteed unique existence of our IPM's MLE. Then, we will discuss how BTM is flawed for an infinite space of $y|x$.

B.1. On IPM's MLE

We make the following mild assumptions about the structure of human preference:

Assumption B.1. $\mathcal{D}_{\cdot|x} = \{y|p(y|x) > 0\}$ is a finite set.

Assumption B.2. $r(x, y) \to -\infty$ for any $y \notin \mathcal{D}_{\cdot|x}$ and $r(x, y) < +\infty$ for any $y \in \mathcal{D}_{\cdot|x}$.

Note that the above two assumptions are just one of many sufficient assumptions that make the partition function exist as a finite real number. Also note that the *finity* of $\mathcal{D}_{\cdot|x}$ and the *infinity* of the space of $y|x$ are two different things that can certainly co-exist. Namely, $\mathcal{D}_{\cdot|x}$ is a subset of the space of $y|x$. The number of y outside of $\mathcal{D}_{\cdot|x}$ is still infinitely large. The two assumptions in plain words are simply that we assume humans will only possibly prefer a finite set of responses. Note that this does *not* mean that the finite set $\mathcal{D}_{\cdot|x}$ cannot be very large.

Definition B.3. The maximum likelihood estimation objective of IPM is the negative log-likelihood of preference data computed as follows:

$$-\Sigma_y^{\infty} p(y|x) \log q_\theta(y|x)$$

where $p(y|x) = \frac{\exp[r(x,y)]}{\Sigma_{y'}^{\infty} \exp[r(x,y')]}$ and $q_\theta(y|x) = \frac{\exp[r_\theta(x,y)]}{\Sigma_{y'}^{\infty} \exp[r_\theta(x,y')]}$.

Given the uniqueness in Corollary A.5, we can argue the following:

Theorem B.4 (Theorem 3.1 in the main content of the paper)**.** *The r_θ that satisfies the slope-1 linearity is the unique minimizer of the IPM's maximum likelihood estimation objective.*

Proof. Again, from the property of Gibb's inequality, we know:

$$\begin{aligned} & q_\theta(y|x) = p(y|x) \\ & \Leftrightarrow \\ & q_\theta(y|x) = \arg\min_q \mathrm{KL}[p(y|x)||q(y|x)] \end{aligned}$$

For the equation on the right, since $p(y|x)$ is a constant w.r.t θ, we can find q_θ is the minimizer of IPM's objective:

$$\begin{aligned} q_\theta(y|x) &= \arg\min_q \mathrm{KL}[p(y|x)||q(y|x)] \\ &= \arg\min_q \mathrm{H}[p(y|x)] - \Sigma_y^{\infty} p(y|x) \log q(y|x) \\ &= \arg\min_q -\Sigma_y^{\infty} p(y|x) \log q(y|x) \end{aligned}$$

We then show that $q_\theta(y|x) = p(y|x)$ is equivalent to slope-1 linearity to complete the proof by taking the logarithm of both sides:

$$\begin{aligned} & r_\theta(x, y) - C_1(x) = r(x, y) - C_2(x) \\ & \Leftrightarrow \\ & r_\theta(x, y) = r(x, y) + C(x) \end{aligned}$$

where $C_1(x) = \log \Sigma_{y'}^{\infty} \exp[r_\theta(x, y')]$, $C_2(x) = \log \Sigma_{y'}^{\infty} \exp[r(x, y')]$ and $C(x) = C_1(x) - C_2(x)$ □

Note that similar proof does not apply to BTM. The fundamental reason is that the $C_1(x)$ and $C_2(x)$ only become constants when there is an infinity in the sum to cancel out all y.

B.2. On Bradley-Terry Model's Flaw

We will show in Proposition B.5 that a very likely choice of $p(y_w, y_l|x)$ will lead to multiple minimizers for the maximum likelihood estimation of BTM. There are also many other choices of $p(y_w, y_l|x)$ are known to cause the existence of multiple minimizers (Ford, 1957; Bong & Rinaldo, 2022), such as when there is no full connectivity of the graph made by pairs from $p(y_w, y_l|x)$, and when all y candidates can only be paired with a single shared winning $\tilde{y}$, etc. Therefore, there is no guarantee for the MLE's uniqueness without imposing additional constraints on $p(y_w, y_l|x)$ (i.e, how the pairs are sampled for DPO). For the loosest *sufficient* constraints discovered so far to ensure the uniqueness, one can refer to Bong & Rinaldo (2022). However, to the best of our knowledge, how such constraints can be applied to DPO has never been studied in the offline alignment literature, which is also out of the scope of this paper. Moreover, in the infinite-candidate scenario, **a constraint that is *both necessary and sufficient* for the uniqueness of BTM's MLE remains unknown** to this day. What makes BTM even more theoretically troublesome in the case of RLHF is that there is also an infinity for the space of x as well. Therefore, strictly speaking, there is an infinite number of BTMs used in DPO. And, the $p(y_w, y_l|x)$ for every x should ensure the uniqueness, in order to make DPO really work as expected. Interestingly, although our EPA loss also needs an infinite number of IPMs in the strict sense, Theorem B.4 (3.1) ensures the MLE uniqueness of all the IPMs.

Proposition B.5. *If there exists a y^* that will never be sampled (i.e., $p(y^*, \cdot|x) = 0$ and $p(\cdot, y^*|x) = 0$), then whenever there is a minimizer for Bradley-Terry's maximum likelihood estimation, it is not unique.*

Proof. Without losing generality, we set $\beta = 1$.

Given the log ratio reward parameterization, we have an intrinsic constraint on r_θ:

$$\sum_{y'}^{\infty} \pi_{\text{ref}} \exp[r_\theta] = \sum_{y'}^{\infty} \pi_\theta = 1$$

If we assume that there is a unique minimizer $r_{\hat{\theta}}$ to BTM's maximum likelihood estimation, it certainly satisfies the above constraint:

$$\sum_{y'}^{\infty} \pi_{\text{ref}} \exp[r_{\hat{\theta}}] = 1$$

We will then show that another reward also follows the constraint (hence a valid log ratio reward) and shares the same expected data likelihood as $r_{\hat{\theta}}$, which contradicts the uniqueness of $r_{\hat{\theta}}$. We define the other reward as:

$$\tilde{r}_{\hat{\theta}}(x,y) = \begin{cases} \log(\exp[r_{\hat{\theta}}(x,y) + A(x)] + \frac{1-exp[A(x)]}{\pi_{\text{ref}}}), & \text{if } y = y^* \\ r_{\hat{\theta}}(x,y) + A(x), & \text{otherwise} \end{cases}$$

where $A(x)$ can be any negative constant w.r.t y. This reward satisfies the constraint because:

$$\begin{aligned}
&\sum_{y'}^{\infty} \pi_{\text{ref}} \exp[\tilde{r}_{\hat{\theta}}(x,y')] \\
&= \sum_{y' \neq y^*}^{\infty} \pi_{\text{ref}} \exp[\tilde{r}_{\hat{\theta}}(x,y')] + \pi_{\text{ref}} \exp[\tilde{r}_{\hat{\theta}}(x,y^*)] \\
&= \sum_{y' \neq y^*}^{\infty} \pi_{\text{ref}} \exp[r_{\hat{\theta}}(x,y') + A(x)] \\
&\quad + \pi_{\text{ref}} \exp[\tilde{r}_{\hat{\theta}}(x,y^*)] \\
&= \exp[A(x)] \sum_{y' \neq y^*}^{\infty} \pi_{\text{ref}} \exp[r_{\hat{\theta}}(x,y')] \\
&\quad + \pi_{\text{ref}} \exp[\tilde{r}_{\hat{\theta}}(x,y^*)] \\
&= \exp[A(x)](1 - \pi_{\text{ref}} \exp[r_{\hat{\theta}}(x,y^*)]) \\
&\quad + \pi_{\text{ref}} \exp[\tilde{r}_{\hat{\theta}}(x,y^*)] \\
&= \exp[A(x)](1 - \pi_{\text{ref}} \exp[r_{\hat{\theta}}(x,y^*)]) \\
&\quad + \pi_{\text{ref}} \exp[r_{\hat{\theta}}(x,y^*) + A(x)] + (1 - \exp[A(x)]) \\
&= \exp[A(x)](1 - \pi_{\text{ref}} \exp[r_{\hat{\theta}}(x,y^*)]) \\
&\quad + \exp[A(x)](\pi_{\text{ref}} \exp[r_{\hat{\theta}}(x,y^*)] - 1) + 1 \\
&= 1
\end{aligned}$$

Note that the constraint makes $\tilde{r}_{\hat{\theta}}$ correspond to a valid policy $\tilde{\pi}_{\hat{\theta}} = \pi_{\text{ref}} \exp[\tilde{r}_{\hat{\theta}}]$ that sums up to 1. The policy is also in the range of $[0, 1]$ because:

1) for $y = y^*$:

$$\begin{aligned}
\tilde{\pi}_{\hat{\theta}} &= \pi_{\text{ref}} \exp[\tilde{r}_{\hat{\theta}}] \\
&= \pi_{\text{ref}} \exp[r_{\hat{\theta}} + A(x)] + 1 - \exp[A(x)] \\
&= \exp[A(x)] \pi_{\text{ref}} \exp[r_{\hat{\theta}}] + 1 - \exp[A(x)] \\
&= \exp[A(x)] \pi_{\text{ref}} \frac{\pi_{\hat{\theta}}}{\pi_{\text{ref}}} + 1 - \exp[A(x)] \\
&= \exp[A(x)](\pi_{\hat{\theta}} - 1) + 1
\end{aligned}$$

and since $A(x) < 0$ and $\pi_{\hat{\theta}} \in [0, 1]$, we have:

$$\begin{aligned}
1 &\geq \exp[A(x)](\pi_{\hat{\theta}} - 1) + 1 \\
&\geq \exp[0](\pi_{\hat{\theta}} - 1) + 1 \\
&= \pi_{\hat{\theta}} \geq 0
\end{aligned}$$

hence $\tilde{\pi}_{\hat{\theta}} \in [0, 1]$;

2) for $y \neq y^*$:

$$\begin{aligned}
\tilde{\pi}_{\hat{\theta}} &= \pi_{\text{ref}} \exp[\tilde{r}_{\hat{\theta}}] \\
&= \pi_{\text{ref}} \exp[r_{\hat{\theta}} + A(x)] \\
&= \exp[A(x)] \pi_{\text{ref}} \exp[r_{\hat{\theta}}] \\
&= \exp[A(x)] \pi_{\text{ref}} \frac{\pi_{\hat{\theta}}}{\pi_{\text{ref}}} \\
&= \exp[A(x)] \pi_{\hat{\theta}} \\
&\in [0, 1]
\end{aligned}$$

Then, we show that this valid log ratio reward is indeed another minimizer because it leads to the same expected likelihood of data as $r_{\hat{\theta}}$.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{p(y_w,y_l|x)} \mathop{\mathbb{E}}_{p(y_w \succ y_l|x)} [\log \sigma(\tilde{r}_{\hat{\theta}}(x,y_w) - \tilde{r}_{\hat{\theta}}(x,y_l))] \\
&= \sum_{y_w,y_l}^{\infty} p(y_w,y_l|x) \mathop{\mathbb{E}}_{p(y_w \succ y_l|x)} [\log \sigma(\tilde{r}_{\hat{\theta}}(x,y_w) - \tilde{r}_{\hat{\theta}}(x,y_l))] \\
&= \sum_{y_w \neq y^*, y_l \neq y^*}^{\infty} p(y_w,y_l|x) \cdot \\
&\quad \mathop{\mathbb{E}}_{p(y_w \succ y_l|x)} [\log \sigma(\tilde{r}_{\hat{\theta}}(x,y_w) - \tilde{r}_{\hat{\theta}}(x,y_l))] \\
&= \sum_{y_w \neq y^*, y_l \neq y^*}^{\infty} p(y_w,y_l|x) \cdot \\
&\quad \mathop{\mathbb{E}}_{p(y_w \succ y_l|x)} [\log \sigma(r_{\hat{\theta}}(x,y_w) + A(x) - r_{\hat{\theta}}(x,y_l) - A(x))] \\
&= \sum_{y_w \neq y^*, y_l \neq y^*}^{\infty} p(y_w,y_l|x) \cdot \\
&\quad \mathop{\mathbb{E}}_{p(y_w \succ y_l|x)} [\log \sigma(r_{\hat{\theta}}(x,y_w) - r_{\hat{\theta}}(x,y_l))] \\
&= \sum_{y_w,y_l}^{\infty} p(y_w,y_l|x) \mathop{\mathbb{E}}_{p(y_w \succ y_l|x)} [\log \sigma(r_{\hat{\theta}}(x,y_w) - r_{\hat{\theta}}(x,y_l))] \\
&= \mathop{\mathbb{E}}_{p(y_w,y_l|x)} \mathop{\mathbb{E}}_{p(y_w \succ y_l|x)} [\log \sigma(r_{\hat{\theta}}(x,y_w) - r_{\hat{\theta}}(x,y_l))]
\end{aligned}$$

Finally, because the possible choices of $A(x)$ are infinite and by assuming the full representation capacity of a large neural network, $\tilde{\pi}_{\hat{\theta}}$ can be represented by some other $\pi_{\tilde{\theta}}$. □

Note that if we are given the $p(y_w, y_l|x)$ defined in Proposition B.5, even if we have an infinite amount of data, we still do not have a unique MLE. This refutes a popular claim that DPO will work better given enough data.

Also, if the rewards of appearing y are shifted by a constant, note that the key intuition of the above proof is that the finiteness of the number of candidates considered in the maximum likelihood estimation loss makes itself invariant as long as some y^* outside of the finite set of candidates never appears. This means a similar property as follows also holds for the Plackett-Luce Model as it also just considers a finite set of candidates.

Corollary B.6. *If there exists a y^* that will never be sampled, then whenever there is a minimizer for Plackett-Luce's maximum likelihood estimation, it is not unique.*

In contrast, a similar proof will not work for our IPM as it considers the entire infinite space of candidates.

B.3. Fitting IPM with Energy Discrepancy

In our paper, since we simply adapt the general theorems given by Schröder et al. (2023) to fit the RLHF context, we only provide proof sketches for quick reference and claim no originality thereof. We encourage readers to refer to the original full proof if more details are needed.

Proof sketch for Theorem 3.2. The first functional derivative of $ED_{x,p(y|x),p(z|y)}[E]$ (we use $E = -r$ to conform to the convention of EBMs) is given by:

$$\begin{aligned}&\frac{d}{d\epsilon} ED_{x,p(y|x),p(z|y)}[E + \epsilon h] \\ &= \mathop{\mathbb{E}}_{x,p(y|x)}[h(x,y)] - \mathop{\mathbb{E}}_{x,p(y|x),p(z|y)} \mathop{\mathbb{E}}_{p_{E,\epsilon}(y'|z)}[h(x,y')]\end{aligned}$$

where

$$p_{E,\epsilon}(y'|z) = \frac{p(z|y') \cdot \exp[-E(x,y') - \epsilon h(x,y')]}{\sum_{y''}^{\infty} p(z|y'') \cdot \exp[-E(x,y'') - \epsilon h(x,y'')]}$$

Then, setting $\epsilon = 0$ and $E = -r_{\text{true}}$, we get the first variation of ED at $E = -r_{\text{true}}$ to be 0 because the second term in the derivative becomes identical to the first term:

$$\begin{aligned}&\mathop{\mathbb{E}}_{x,p(y|x),p(z|y)} \mathop{\mathbb{E}}_{p_{E=-r_{\text{true}},\epsilon=0}(y'|z)}[h(x,y')] \\ &= \mathop{\mathbb{E}}_{x,p(y|x),p(z|y)} \sum_{y'}^{\infty} \frac{p(z|y') \cdot \exp[r_{\text{true}}(x,y')]}{\sum_{y''}^{\infty} p(z|y'') \cdot \exp[r_{\text{true}}(x,y'')]} \cdot h(x,y') \\ &= \mathop{\mathbb{E}}_{x,p(y|x),p(z|y)} \sum_{y'}^{\infty} \frac{p(z|y')p(y'|x)}{\sum_{y''}^{\infty} p(z|y'')p(y''|x)} \cdot h(x,y') \\ &= \mathbb{E}_x \sum_{y}^{\infty} \sum_{z} p(y|x)p(z|y) \frac{\sum_{y'}^{\infty} p(z|y')p(y'|x)h(x,y')}{\sum_{y''}^{\infty} p(z|y'')p(y''|x)} \\ &= \mathbb{E}_x \sum_{y'}^{\infty} p(y'|x)h(x,y') \sum_{y}^{\infty} \sum_{z} \frac{p(y|x)p(z|y)p(z|y')}{\sum_{y''}^{\infty} p(z|y'')p(y''|x)} \\ &= \mathbb{E}_x \sum_{y'}^{\infty} p(y'|x)h(x,y') \sum_{z} p(z|y') \frac{\sum_{y}^{\infty} p(z|y)p(y|x)}{\sum_{y''}^{\infty} p(z|y'')p(y''|x)} \\ &= \mathbb{E}_x \sum_{y'}^{\infty} p(y'|x)h(x,y') \sum_{z} p(z|y') \\ &= \mathbb{E}_x \sum_{y'}^{\infty} p(y'|x)h(x,y') \\ &= \mathbb{E}_{x,p(y|x)} h(x,y)\end{aligned}$$

With the first variation being 0, we will then only need to show that the second variation of ED at $E = -r_{\text{true}}$ is strictly positive to complete the proof that $E = -r_{\text{true}}$ is the global unique minimizer of ED. This can be done by showing that the second derivative becomes an expectation of $\text{Var}_{p(y|z)}[h(y)]$ which cannot be negative or zero because we assume $\text{Var}_{p(y|z)}[y] > 0$ (i.e., $\text{Var}[Y|Z] > 0$). Concretely, the second derivative of ED is given by:

$$\begin{aligned}&\frac{d^2}{d\epsilon^2} ED_{x,p(y|x),p(z|y)}[E + \epsilon h] \\ &= - \mathop{\mathbb{E}}_{x,p(y|x),p(z|y)} \frac{d^2}{d\epsilon^2} \mathop{\mathbb{E}}_{p_{E,\epsilon}(y'|z)}[h(x,y')] \\ &= \text{(see (Schröder et al., 2023)'s Lemma 2)} \\ &\quad \mathop{\mathbb{E}}_{x,p(y|x),p(z|y)} \left[\mathop{\mathbb{E}}_{p_{E,\epsilon}(y'|z)}[h^2(x,y')] - \left(\mathop{\mathbb{E}}_{p_{E,\epsilon}(y'|z)}[h(x,y')] \right)^2 \right] \\ &= \mathop{\mathbb{E}}_{x,p(y|x),p(z|y)} [\text{Var}_{p_{E,\epsilon}(y'|z)}[h(x,y')]]\end{aligned}$$

Setting $\epsilon = 0$ and $E = -r_{\text{true}}$, we will get the positive second variation of ED at $E = -r_{\text{true}}$. □

Proof sketch for Theorem 3.3. Given the property that

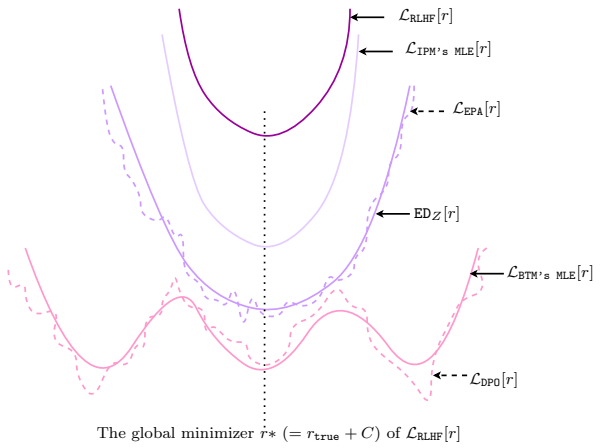

Figure 4. A simplified illustration of the topology of the functionals mentioned in the paper's theorems. The vertical axis represents the value of each functional. The horizontal axis represents the space of r.

$\Sigma_y p(z|y)f(y) = \Sigma_y p(y|z)f(y)$, the $\Sigma_{y'} p(z|y')$ in the definition of ED becomes an expectation over y, i.e., $\mathbb{E}_{p(y'|z)}$. This enables a statistic estimate using the normal and a modified Strong Law of Large Numbers (Majerek et al., 2005) for the expectations outside and inside of the logarithm, respectively. □

B.4. The Perspective of Functional Analysis

An elegant way to understand the theoretical advantage of EBM/EPA over BTM/DPO is through the lens of functional analysis. As illustrated in Figure 4, our EBM's maximum likelihood estimation loss, the energy discrepancy, and the RLHF loss can be regarded as three functionals sharing the same unique global minimizer. The EPA loss, being a statistical estimate[4] of the energy discrepancy, is yet another functional. However, the Strong Law of Large Numbers ensures that the error between it and the energy discrepancy will almost surely vanish with enough amount of negatives drawn from a properly designed data perturbing source Z. Therefore, one can view the EPA loss as a locally fuzzy approximation of the energy discrepancy. The maximum likelihood estimation loss of BTM is a functional that may have multiple minimizers. Therefore, without explicit constraints to prevent this from happening, chances are that optimizing the DPO loss (which is strictly speaking also just a statistical estimate of the maximum likelihood of data, probably with less fuzziness) will lead to a different solution than the minimizer of the RLHF loss. Increasing the amount of data for DPO can certainly mitigate the fuzziness but can do nothing to avoid the undesirable structure of BTM's maximum likelihood estimation loss governed by $p(y_w, y_l|x)$.

[4] Strictly speaking, there is also a constant shift $\log M$, which we removed from the energy discrepancy to derive Eq.(10) because it has no topological impact in terms of optimization problems.

Table 6. Preliminary results on tuning β for IPO.

β	MT-BENCH (EPOCH #1 / #2 / #3)
0.1	6.73 / 6.88 / 6.87
0.01	7.20 / 7.31 / 7.23

C. More Experimental Details

C.1. Implementation Details

C.1.1. More Hyperparameter Details

We use 8 A100/A800 GPUs (80G Memory) with ZeRO3 parallelism to train each model in this paper. Global batch size is fixed to 64. For experiments in Figure 2, we run two rounds with different seeds (0 and 1) for each model configuration. Other experiments are only conducted with seed 0. The reason we set β to 0.01 in the main experiments is two-fold.

- Firstly, results (see Figure 2.(a)) of tuning β for multiple values ranging from 0.01 to 0.5 show that β close to 0.01 (low-KL region) will make both EPA and DPO achieve higher rewards.

- Secondly, $\beta = 0.01$ is also the default setting for NCA and infoNCA (Chen et al., 2024a) and also a recommended setting for Mistral-based DPO (Tunstall et al., 2023) and KTO (Ethayarajh et al., 2024). For IPO's best β, we did preliminary experiments with 0.01 and 0.1, and found 0.01 works better (see Table 6).

C.1.2. Loss Modifications Tricks

In Table 7, we list how the tricks are applied to the DPO loss or the narrow EPA loss. For more detailed properties of them, one can refer to the corresponding previous work. In summary, length penalty (Park et al., 2024), length normalization (Yuan et al., 2023; Hong et al., 2024; Meng et al., 2024), constant or dynamic margins (Meng et al., 2024; Amini et al., 2024), removal of the reference model (Xu et al., 2024; Hong et al., 2024; Meng et al., 2024), on-policy weighting (Zhou et al., 2024) and addition of SFT loss (Xu et al., 2024; Hong et al., 2024), etc.

Table 7. Notation and function of loss modification tricks proposed in the offline alignment literature. We consider on-policy weighting ($+w_{op}$), length penalty ($+len_p$), length normalization ($+len_n$), constant margin ($+m_c$), removal of the reference model ($-ref$), and addition of a SFT loss ($+\mathcal{L}_{sft}$).

NOTATION	TRICK FUNCTION
$-ref$	$\log \frac{\pi_\theta(y\|x)}{\pi_{\text{REF}}(y\|x)} \to \log \pi_\theta(y\|x)$
$+\mathcal{L}_{sft}$	$\mathcal{L} \to \mathcal{L} - \log \pi_\theta(y_w\|x)$
$+len_p$	$r_\theta(y\|x) \to r_\theta(y\|x) - \alpha\|y\|$
$+len_n$	$r_\theta(y\|x) \to \frac{1}{\|y\|} r_\theta(y\|x)$
$+m_c$	$r_\theta(y_l\|x) \to r_\theta(y_l\|x) + m_c$
$+w_{op}$	$\mathcal{L} \to \Pi_{*\in\{w,l\}} (\Pi_t \frac{\pi_\theta(y_*^t \| y_*^{0:t-1}, x)}{\Sigma_{y' \in Voc} \pi_\theta^2(y' \| y_*^{0:t-1}, x)})^{\frac{1}{\|y\|}} \cdot \mathcal{L}$

C.1.3. COMPARING EPA WITH OTHER MULTI-RESPONSE LOSSES

It appears that EPA and two multi-response losses (InfoNCA and DPO-PL) are very similar because they are all contrastive losses in the form of negative log softmax. However, there is a clear distinction between EPA and the other losses in terms of the required information in the training data and the samples being contrasted (exemplified by Table 8). This is a direct consequence of using different probabilistic models. EPA comes from our IPM. DPO-PL comes from the Plackett-Luce Model, and InfoNCA comes from an energy-based model for the distribution of the optimal policy.

C.2. DPO vs. EPA from More Perspectives

C.2.1. ALIGNMENT TAX

Preference alignment is usually associated with an alignment tax: forgetting certain capabilities (e.g., math problem solving) while enhancing others (e.g., safety, truthfulness, and helpfulness). We acknowledge that EPA might exhibit a higher alignment tax than DPO due to the results in Table 9, where we report metrics on GSM8k (Cobbe et al., 2021), MMLU (Hendrycks et al., 2020), and Winograd (Tikhonov & Ryabinin, 2021). However, a more comprehensive future work on this issue is still necessary for a reliable conclusion. For example, although GSM8k and MMLU are of relatively higher correlation with human preference than other Exact-Match-based benchmarks, they have poorer correlations than the ones used in the main paper (MT-Bench and Alpaca-Eval 2.0) (Dubois et al., 2024). Therefore, a more aggressive alignment method with human preference could cause the lower scores on GSM8k and MMLU.

C.2.2. PROBING THE SLOPE-1 LINEARITY

Reward models in the classical sense only need to satisfy ranking consistency with the corresponding true rewards. However, for ideal offline optimization of the RLHF loss, a learned log ratio reward and its corresponding true reward have to satisfy the slope-1 linearity, a much stronger requirement than ranking consistency. The reason is that, to satisfy ranking consistency, the relationship between the two rewards can be any monotonically increasing function, including those that are *non-linear* or *linear with the slope being any positive number*.

Therefore, metrics that are based on ranking consistency to evaluate the log ratio reward (i.e., the learned reward model) are not meaningful when it comes to offline alignment. Instead, we evaluate the log ratio reward by probing how well the linearity is approximated. For this purpose, we need multiple samples of y (responses) given an x (prompt) and their true rewards. The test split of the Ultrafeedback data (Tunstall et al., 2023) can fulfill this purpose because there are four y for each x and they are scored using the same scoring scheme used for our training data (i.e., UF-all and UF-binarized). We randomly sample 500 prompts from the test split to speed up evaluation while preserving the general reliability.

Firstly, we consider only how linear the relationship between the two rewards is, regardless of the slope. This is exactly the essence of Pearson correlation analysis. We compute the Pearson coefficient between the two rewards over the four y for each prompt. Then, we compute the mean over all 500 prompts as a metric. We report the results in the first column of Table 1 (in the main content of the paper).

Secondly, we study both the "slope-1" and the "linearity". We do this via linear regression with the slope fixed to 1. Specifically, given a prompt, we need to fit a linear regression model $r_{\text{learned}} = 1 \cdot r_{\text{true}} + b$ to the 4 coordinates of $(r_{\text{learned}}, r_{\text{true}})$. With simple algebra, the optimal value of b that minimize the linear regression error $\epsilon = \Sigma|r_{\text{true}} - r_{\text{learned}} + b|^2$ can be analytically expressed as $\hat{b} = (1/4) \cdot \Sigma(r_{\text{learned}} - r_{\text{true}})$. Thus, we use $\hat{b}$ to compute the minimal error for each given prompt, and then compute the average minimal error $\hat{\epsilon}$ over all 500 prompts as the metric (see the second column of Table 1). We can also shift the four r_{learned} for each prompt by the constant $-\hat{b}$, which should move all regression lines for the slope-1 linearity to the same location: the diagonal $r_{\text{learned}} = 1 \cdot r_{\text{true}}$. This allows us to visualize the overall degree of how well the slope-1 linearity is approximated. As shown in Figure 5, we can see that for the top 10% prompts with the best $\hat{\epsilon}$ (i.e., smallest $\hat{\epsilon}$), the slope-1 linearity is well approximated for both EPA and DPO. However, we can observe that DPO is slightly off

Table 8. Comparing EPA, InfoNCA and DPO-PL in the setting of 3 responses per prompt (omitting x for clarity).

NAME	DATA	LOSS
EPA	$(y_w, y_{l_1}, y_{l_2}), (y_w \succ y_{l_1}), (y_w \succ y_{l_2})$	$-\log \frac{\exp(r_\theta(y_w))}{\exp(r_\theta(y_w))+\exp(r_\theta(y_{l_1}))+\exp(r_\theta(y_{l_2}))+\exp(r_\theta(y_{wk_1}))+\exp(r_\theta(y_{wk_2}))}$
DPO-PL	$(y_w, y_{l_1}, y_{l_2}), (y_w \succ y_{l_1} \succ y_{l_2})$	$-\log \frac{\exp(r_\theta(y_w))}{\exp(r_\theta(y_w))+\exp(r_\theta(y_{l_1}))+\exp(r_\theta(y_{l_2}))} - \log \frac{\exp(r_\theta(y_{l_1}))}{\exp(r_\theta(y_{l_1}))+\exp(r_\theta(y_{l_2}))}$
INFONCA	$(y_w, y_{l_1}, y_{l_2}), (y_w \succ y_{l_1}), (y_w \succ y_{l_2})$	$-\log \frac{\exp(r_\theta(y_w))}{\exp(r_\theta(y_w))+\exp(r_\theta(y_{l_1}))+\exp(r_\theta(y_{l_2}))}$

Table 9. EPA has slightly higher alignment tax than DPO.

METHOD	GSM8K-5SHOT	MMLU	WINOGRAD
SFT	0.421 (0.014)	**0.598** (0.004)	**0.800** (0.006)
+DPO	**0.463** (0.014)	0.593 (0.004)	0.790 (0.006)
+EPA	0.419 (0.014)	0.591 (0.004)	0.790 (0.006)

from the slope-1 linearity for the medium 10% and much so for the worst 10%. On the other hand, for EPA, although the points are also gradually spreading out when we move towards the worst 10%, they are still distributed along the direction of the diagonal. This phenomenon means that EPA is closer to the slope-1 linearity than DPO, especially for the worst group of prompts.

Table 10. EPA vs DPO applied to Llama3-8B based SFT model.

METHOD	MT-BENCH (EPOCH #1 / #2 / #3)
DPO	6.94 / 6.83 / 6.71
EPA	6.84 / 6.98 / **7.04**

Table 11. EPA vs DPO on Arena Hard.

TRAINING DATA	METHOD	ARENA HARD (%)
UF-BINARIZED	DPO	12.0
	EPA	**16.3**
UF-ALL	DPO-PL	13.0
	EPA	**16.9**

C.2.3. GENERALITY OF THE PROPOSAL

Although we believe the experimental configuration described in the main content of the paper provides sufficient support for our core theoretical argument, we present additional results here using a different base model and training dataset. Specifically, we train a Llama3-8B based SFT model[5] on a cleansed version (Álvaro Bartolomé Del Canto et al., 2024) of the widely used but small dataset: Intel/orca_dpo_pairs. As shown in Table 10, EPA is still better performing than DPO. In addition, we also report the best checkpoints for EPA and DPO/DPO-PL on Arena Hard (Li et al., 2024), yet another benchmark that is considered more difficult. The result is consistent with Alpaca Eval 2.0 and MT-Bench as reported in Table 2, further validating the superiority of EPA.

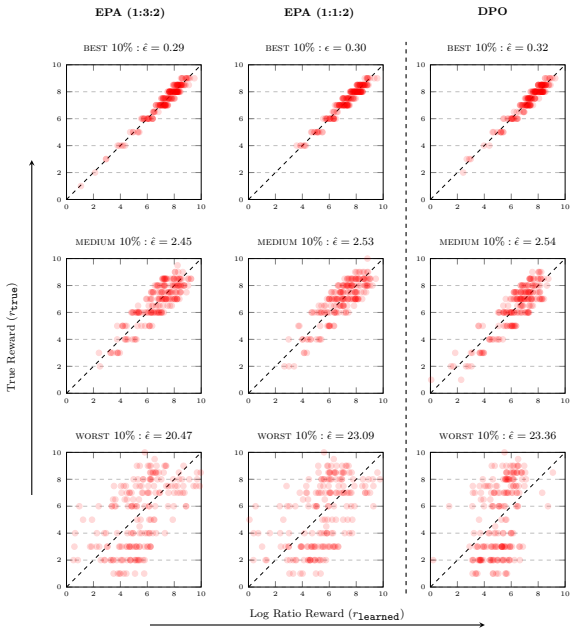

Figure 5. EPA vs DPO in terms of the slope-1 linearity. Each average slope-1 linear regression error $\hat{\epsilon}$ is computed on $500 \times 10\% = 50$ prompts. Although the difference between EPA and DPO only becomes noticeable for the "WORST 10%" group in the visualization (i.e., how close the partially transparent red dots are to the diagonal), the difference in $\hat{\epsilon}$ is conspicuous.

[5] https://huggingface.co/princeton-nlp/Llama-3-Base-8B-SFT