DATA-CENTRIC HUMAN PREFERENCE OPTIMIZATION WITH RATIONALES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning from human feedback plays a crucial role in aligning language models towards human preferences, traditionally represented through comparisons between pairs or sets of responses within a given context. While many studies have enhanced algorithmic techniques to optimize learning from such data, this work shifts focus to improving preference learning through a data-centric approach. Specifically, we propose enriching existing preference datasets with machine-generated rationales that explain the reasons behind choices. We develop a simple and principled framework to augment current preference learning methods with *rationale* information. Our comprehensive analysis highlights how rationales enhance learning efficiency. Extensive experiments reveal that rationale-enriched preference learning offers multiple advantages: it improves annotation efficiency, accelerates convergence to higher-performing models, and reduces verbosity bias and hallucination. Furthermore, this framework is versatile enough to integrate with various preference optimization algorithms. Overall, our findings highlight the potential of re-imagining data design for preference learning, demonstrating that even freely available machine-generated rationales can significantly boost performance across multiple dimensions.

1 INTRODUCTION

Preference tuning is an important step in the language model training process so as to productize and deploy them, whose goal is to align the model towards human preferences and prevent the model from unwanted behavior (Christiano et al., 2017; Stiennon et al., 2020; Bakker et al., 2022).

These preferences are typically introduced into the dataset through prompts and ranked responses. This dataset is then utilized by reinforcement learning from human feedback (RLHF) methods (Ouyang et al., 2022) to optimize reward or preference models, thereby aligning the language model. (Schulman et al., 2017) proposes a reinforcement learning-based algorithm to train the model by maximizing the reward given the reward function, where the first step involves learning a reward model to replicate human preferences. Alternatively, (Rafailov et al., 2024) proposes a direct preference optimization (DPO) algorithm, which avoids training a separate reward model and optimizes the policy through implicit Bradley-Terry reward modeling with a single objective.

However, these methods often face several challenges such as overfitting Azar et al. (2024), performance degradation Pal et al. (2024), reward exploitation Amodei et al. (2016), or the generation of excessively long inputs Park et al. (2024). In addition, collecting preference datasets can be costly Tan et al. (2024), but without sufficient samples, these methods would risk underfitting Jinnai & Honda (2024). Various studies aim to address these issues through improved algorithmic designs, either by regularizing the objective Pal et al. (2024); Amini et al. (2024); Park et al. (2024) or by introducing new formulations Ethayarajh et al. (2024); Hong et al. (2024); Yuan et al. (2023); Munos et al. (2023); Swamy et al. (2024); Wu et al. (2024).

In our study, we transition from an algorithmic to a data-centric approach, posing the key question: *How can enhancing the preference dataset aid the model in boosting its performance and efficiency in preference learning?* By rethinking preference learning through the lens of data, we aim to discover new insights and opportunities that can unlock the potential for more robust, data-efficient preference learning. Given the current setup of preference datasets, an important question arises: why would a certain response be preferred over another? For obvious cases, it is simple for humans to understand

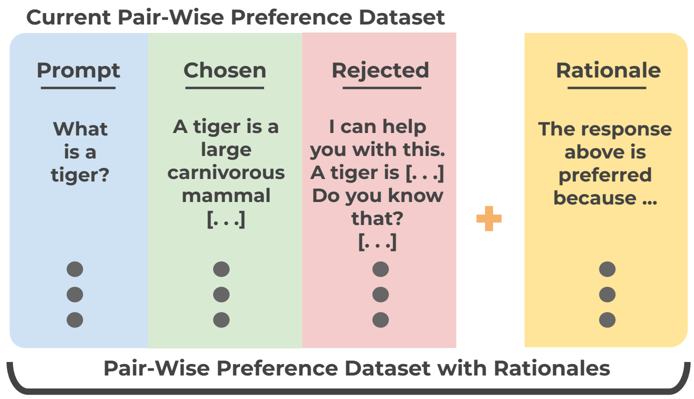

Figure 1: Comparison between the current pair-wise preference dataset used for preference learning and the enriched dataset with added rationales.

the preference. However, when the responses are closely matched, understanding the preference without any explanation becomes challenging. Even superficial features such as length might not serve as a straightforward metric to determine preference. In one instance, a longer response may be favored for its comprehensiveness, while in another instance, a shorter response might be preferred for its conciseness. Another consideration regarding preferences is that individuals might have varying preferences for different reasons, and without explicitly outlining these reasons, one would be unable to discern the underlying rationale. Given these challenges, the model will struggle to learn these preferences without any explanations, causing data inefficiency, and worse, it could learn the wrong cues, decreasing performance.

For the reasons outlined above, we propose a natural extension to the current dataset structure by enriching the preferences with rationales, aiding the model in better understanding during preference learning. Rationales explain why one response is preferred over another for a given prompt. This idea also draws inspiration from social studies (Mitchell et al., 1986; Chi et al., 1994; Crowley & Siegler, 1999; Williams & Lombrozo, 2010), showing that adding explanations to answers improves one's understanding of the problem compared to individuals who do not provide explanations.

Contributions. This paper provides a new data-centric perspective on preference learning. We list the summary of contributions:

- We introduce rationales into the human preference learning framework, where rationales explain the reasons behind the preference for a particular response. In practice, these rationales can be generated in a cost-effective manner by prompting an off-the-shelf language model, which may or may not have undergone preference learning.
- We derive a straightforward formulation for the preference function to extend the rationales and show how to adapt our method to current preference learning algorithms such as DPO.
- We analytically examine the impact of the rationale on preference training through the lens of information theory. Our theoretical analysis demonstrates that highly informative rationales can improve preference prediction accuracy and reduce sample complexity.
- We empirically show the impact of preference learning with rationales, highlighting improvements in both performance and annotation efficiency compared to baseline methods. Specifically, the rationale-enriched DPO model can save up to $3\times$ the annotated data required by the vanilla DPO model. With the same amount of data, it can improve the winrate against the supervised fine-tuned model by $8-9\%$. Further, the rationale-based DPO model shows reduced susceptibility to verbosity bias and truthfulness degradation compared to DPO. We demonstrate the flexibility and effectiveness of our approach by extending rationales to ORPO.
- We showcase the efficacy of rationales generated by the off-the-shelf models with $\leq 8B$ parameters on the preference learning. We emphasize the importance of high-quality data for improved preference learning.
- We release our code and datasets to facilitate further research in this direction.

In a broader context, our approach presents a new paradigm for data-centric research in language modeling: rather than focusing on pruning samples to distill the most informative pieces from a dataset Albalak et al. (2024), we explore how to enrich each sample's information content and examine its impact. The promising results presented in this paper demonstrate the effectiveness of enhancing individual samples' information content in preference learning and suggest that this approach may hold potential for improving learning in other domains.

2 Related Work

RLHF with Reward Modeling. Tuning large language models to align their outputs towards human preferences is crucial for controlling model behavior and maintaining desirable boundaries (Casper et al., 2023). To achieve this, RLHF has been introduced; aligning models through preference training (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020). Schulman et al. (2017) describe a method that typically involves two stages. The first stage learns a reward model using a preference dataset often modeled under the Bradley-Terry model Bradley & Terry (1952). The second stage fine-tunes the target policy model to maximize the rewards from the reward model, employing algorithms such as proximal policy optimization (PPO) proposed by Schulman et al. (2017) and adopted in Ouyang et al. (2022). A direct preference optimization (DPO) method that implicitly models the reward function was introduced by Rafailov et al. (2024). However, Azar et al. (2024) observe that RLHF and DPO are prone to overfitting due to the assumptions of the Bradley-Terry model. Conversely, Pal et al. (2024) explore the possibility of DPO underfitting when dealing with challenging responses that are difficult for the model to distinguish. Additionally, Park et al. (2024) note that DPO can exploit response length to maximize reward, proposing a length-regularized DPO (R-DPO) to address this issue. This should not be confused with our rationale-based DPO (RDPO) method. Interestingly, we observe that if rationales mention conciseness as a feature, then the length of responses is significantly reduced compared to SFT and DPO responses. The learning dynamics during preference tuning are analyzed in Im & Li (2024), emphasizing the importance of high-quality preference datasets for effective learning. They find that the more distinguishable the response pairs, the easier it is for the model to learn, leading to faster convergence. This has also been observed in Pal et al. (2024). However, designing such datasets is challenging, and it also remains important for models to learn from such nuanced responses, which appear in practice. We try to address the difficulty of model learning from intricate preferences by providing rationales during preference training. To improve efficiency over DPO, which requires an intermediate step to train the reference model, odds ratio preference optimization (ORPO) was introduced by Hong et al. (2024) to eliminate this step. Another method by Ethayarajh et al. (2024) adapts the Kahneman-Tversky human utility model to handle preference datasets with a single response (either chosen or rejected), removing the need for training the model on both responses. Conversely, Yuan et al. (2023) propose a preference method that considers multiple ranked responses for a prompt and optimizes over them. Our method can complement these methods by adding rationales into training. In this paper, we demonstrate an extension of our framework to ORPO.

General Preference Modeling. Reward modeling, however, can incentivize undesirable behaviors, such as "reward hacking" (Amodei et al., 2016), where agents maximize rewards without achieving the desired objective. Overfitting is another challenge, as exemplified in Azar et al. (2024). While effective for comparing two responses, the Bradley-Terry preference modeling relies on the assumption of transitivity, which may not hold true in practice (Bertrand et al., 2023). To address this, Munos et al. (2023) introduced general preference modeling, which directly learns general preferences by formulating a two-player, constant-sum game between policies. The goal is to maximize the probability of generating the preferred response against the opponent. The solution is the Nash equilibrium of this game, where payoffs are derived from the general preference function. Building upon this work, Munos et al. (2023) proposed an algorithm for the regularized general preference model, while Swamy et al. (2024) developed a solution for the unregularized formulation and introduced self-play preference optimization (SPO) as an iterative algorithm to reach the optimal solution. However, SPO suffers from data inefficiency due to its two-timescale update rules. To address this, Rosset et al. (2024) introduced an efficient direct Nash optimization (DNO) method that leverages the DPO formulation in practice. Additionally, Wu et al. (2024) proposed an efficient, scalable, iterative self-play method that generates responses generally preferred over others.

While previous efforts have introduced algorithmic enhancements for preference tuning, they have been limited to the existing framework of preference datasets with prompts and ranked responses. In contrast, our work is first to introduce rationales, a data-centric solution, into preference learning.

Learning with Rationales. The supervised learning framework typically involves training a model to learn the ground truth label for a given prompt without providing explicit explanations for the associations, which can lead to the model learning incorrect cues. To mitigate this issue, rationales have been integrated into the framework, offering explanations for the given associations. These rationales initially were generated by humans (Zaidan et al., 2007; Ross et al., 2017; Hase & Bansal, 2021; Pruthi et al., 2022). However, due to the high cost of human labor and the development of more capable large language models, rationales are now often automatically generated by these models, reducing the need for human involvement (Wei et al., 2022; Kojima et al., 2022). Given rationales, they have been used as guiding aids by incorporating them directly into the prompt during the training phase (Rajani et al., 2019; Zelikman et al., 2022; Huang et al., 2023) or at the inference stage (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022). Besides using them as additional context within the prompt, rationales can also serve as labels to train models to generate such explanations for their predictions (Wiegreffe et al., 2021; Narang et al., 2020; Eisenstein et al.; Wang et al.; Ho et al., 2023; Magister et al., 2022; Li et al., 2023a). In similar manner, rationales have been applied in knowledge distillation, where they are generated by a more capable models to supervise weaker models (Hsieh et al., 2023; Chen et al., 2024). In parallel with these advancements, we introduce rationales into the preference learning landscape, where rationales are used to explain the preference of one answer over another. Our findings demonstrate the effectiveness of rationales in preference learning, even when generated by the same model or a smaller-sized model.

3 METHOD

In this section, we introduce the incorporation of rationales into preference learning and show the derivation of adapting current methods. We present a demonstration of extending the direct preference optimization (DPO) algorithm to incorporate the rationales, while similar extensions can be applied to other variants of DPO. Further, we analyze theoretically the possible impact of rationales through the perspective of information theory.

3.1 PRELIMINARIES

Notations. Let $\mathcal{D}$ denote the pair-wise preference dataset of size N, $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$, where $x^{(i)}$ is a context, $y_w^{(i)}$ is the preferred/chosen/winning response to the context $x^{(i)}$ over the unpreferred/rejected/losing response $y_l^{(i)}$. Let π_θ and π_{ref} denote the policy to be preference optimized and the reference policy respectively. In our setting, the policies are the language model to be preference trained and the base or supervised fine-tuned SFT model, respectively. To compute the joint probability of the autoregressive language model π generating the response y given the prompt x, we compute the product of probabilities after observing each token: $\pi(y|x) = \Pi_{t=0}^{|y|}\pi(y_t|x, y_{0:t})$.

Reward Modeling with DPO. In the RLHF process (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Bai et al., 2022; Ouyang et al., 2022; Rafailov et al., 2024), the goal is to align the language model towards human preferences. The preferences ranking from the dataset $\mathcal{D}$ is assumed to be sampled from the latent reward function $r^*(x, y)$ and the preference function is assumed to be generated by the Bradley-Terry model (Bradley & Terry, 1952): $p^*(y_w \succ y_l|x) = \sigma(r^*(x, y_w) - r^*(x, y_l))$, where σ is the sigmoid function. The reward function then can be estimated by minimizing the log-likelihood of the following objective $\mathcal{L}(r, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma(r(x, y_w) - r(x, y_l))\right]$. Then the next step is to tune the language model with the reward model as follows by maximizing the rewards and not diverging from the fixed reference model: $\max_{\pi_\theta}\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left[r(x, y)\right] - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta(y|x)\|\pi_{\text{ref}}(y|x)\right]$, where β is a hyperparameter measuring the divergence between two policies. Alternatively, with a reparametrization of the Bradley-Terry preference model (Rafailov et al., 2024), the preference function can be expressed in terms of policy π^*:

$$p^*(y_w \succ y_l|x) = \sigma\left(\beta\log\frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi^*(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right). \tag{1}$$

Thus, to estimate the policy, Rafailov et al. (2024) proposes to directly minimize the log-likelihood of the following DPO loss: $\mathcal{L}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$.

3.2 Formulation of Preference Learning with Rationales

While preferences are modeled given the preferred and unpreferred responses, there are nuances in the responses that are obscure for the model to comprehend and catch the differences between them. Therefore, our goal is to help the model learn the preferences by providing guidance cues in the preference tuning process, which we call *rationales*. Rationales explain why a given response is preferred over the other response. For that reason, we extend the current preference learning with a data-centric technique to incorporate rationales and we term this the *rationale-enriched preference function*, where the updated preference function is formulated as $p^*(y_w \succ y_l, r|x)$ and r is the rationale from the updated dataset $\mathcal{D}' = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}, r^{(i)}\}_{i=1}^N$. By the chain rule, we arrive at:

$$p^*(y_w \succ y_l, r|x) = p^*(y_w \succ y_l|x) \cdot p^*(r|x, y_w \succ y_l), \tag{2}$$

where the first term is the pair-wise preference term modeled in Section 3.1, and the second term is the probability of the rationale r given the context x and the preference $y_w \succ y_l$. Given the policy π^*, we can retrieve the probability of generating the rationale r given the context x and the preference $y_w \succ y_l$, $\pi^*(r|x, y_w \succ y_l)$. Similarly when retrieving the probability of generating responses y_w and y_l for the prompt x, which are given in the preference dataset $\mathcal{D}'$, we can also retrieve the probability of generating rationale r given x,y_w, and y_l, where $(x, y_w, y_l, r) \sim \mathcal{D}'$. In practice, we ask the policy language model to explain why the response y_w is preferred over the response y_l for the prompt x and retrieve the probability of generating the rationale r. Thus, $p^*(r|x, y_w \succ y_l) = \pi^*(r|x, y_w \succ y_l)$.

Adaptation to DPO Loss. After deriving the rationale-enriched preference learning function, we extend the DPO method to incorporate rationales. By substituting $p^*(y_w \succ y_l|x)$ from Equation 1 and $p^*(r|x, y_w \succ y_l)$ into Equation 2, we can express the rationale-enriched preference function in terms of an optimal policy π^*:

$$p^*(y_w \succ y_l, r|x) = \sigma\left(\beta \log \frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi^*(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right) \pi^*(r|x, y_w \succ y_l). \tag{3}$$

We can optimize our policy π_θ through maximum likelihood using the following objective over the updated preference dataset $\mathcal{D}' = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}, r^{(i)}\}_{i=1}^N$, which we term as rationale-DPO (RDPO):

$$\begin{aligned}\mathcal{L}_{\text{RDPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l,r)\sim\mathcal{D}'}\Bigg[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right) + \\ \gamma \log \pi_\theta(r|x, y_w \succ y_l)\Bigg],\end{aligned} \tag{4}$$

where γ is the added hyperparameter for weighting the impact of rationales on the loss.

4 Evaluation

In this section, we evaluate the impact of rationales on preference learning. We conduct multiple experiments with two main goals in mind: **(1)** to understand how the added rationales affect the efficacy and efficiency of current preference learning algorithms, and **(2)** to determine the significance of rationale quality for effective learning.

4.1 Experimental Setup

Datasets. For our analysis, we focus on two popular preference datasets: Orca DPO Pairs (Intel, 2024), which is a pairwise preference dataset version of Orca (Mukherjee et al., 2023), and a binarized UltraFeedback (Tunstall et al., 2023), which is a pair-wise version of UltraFeedack (Cui et al., 2023). For each dataset, we take 12,000 samples as training set and 512 fixed samples as test set for winrate evaluations. We generate rationales and add rationales to the current datasets. We refer readers to Appendix B.5 for details on generating rationales. For evaluating hallucination, we adopt the TriviaQA dataset (Joshi et al., 2017) and use LM Evaluation Harness (Gao et al., 2023) code to measure the exact match (EM) accuracy on the dataset. Given the test sets, we sample the responses from models trained with preference learning methods and compare the performance between the models by measuring the winrates between the corresponding responses.

Models. We investigate preference training on various large language models: Mistral-7B-v0.1, Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Zephyr-7B-Beta(Tunstall et al., 2023), and Llama3-8B-Instruct (AI@Meta, 2024). We use GPT-4o (Achiam et al., 2023) as a judge to evaluate the responses generated by the models and to retrieve the winrate scores. While in this section, we mainly study the Mistral-7B-Instruct-v0.2 model (unless explicitly specified) with rationales generated by this model, we also provide full results on remaining models with ablation on hyperparameters in Appendix B.2.

Methods. In our experiments, we study the integration of rationales into preference learning frameworks, such as DPO (Rafailov et al., 2024), which requires the SFT model for the reference model, and ORPO (Hong et al., 2024), which does not. To ensure fair comparison between DPO and RDPO, we fine-tune the base model with supervised fine-tuning (SFT) only using the chosen responses from the preference dataset for a single epoch. We extend the code implementation from human-aware loss functions (HALOs) repository (Ethayarajh et al., 2023) to adapt to our methodology and borrow the hyperparameters for each of the above methods in our study.

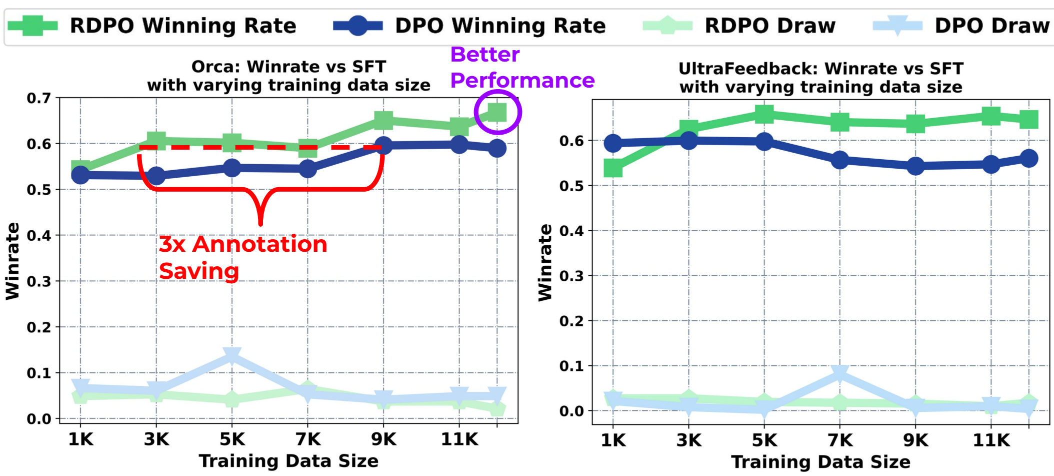

Figure 2: Winrare comparison between the models trained with RDPO and DPO. **Left:** Winrate against the SFT model trained on the Orca dataset. **Right:** Winrate against the SFT model trained on the Ultrafeedback dataset. X-axis denotes the training data size used for preference training of DPO and RDPO models.

4.2 Performance of Preference Learning with Rationales

Versus SFT. We examine how adding rationales to the current preference learning algorithms can impact performance. We compare the responses generated by the preference-aligned model against the responses generated by the SFT model and measure the winrates scores using the GPT-4o evaluator. To study data annotation efficiency, we train the models on various training data sizes, ranging from $1{,}000$ to $12{,}000$ data points, for both RDPO and DPO training. We observe on the left side of Figure 2 that both models, DPO and RDPO models, achieve a better winning rate over the SFT model (more than 50%) on the Orca dataset with an increasing winning trend when training data increases. Additionally, we observe as the DPO model converges to around 60% winning rate at the $9{,}000$ mark against the SFT model, the RDPO model achieves this rate at a smaller training data size with $3{,}000$ training data points. Furthermore, we observe that the RDPO model can reach an even higher winning rate against the SFT than the DPO model, reaching above the 66% winning rate. We see a similar observation with the models trained on the UltraFeedback dataset on the right side of Figure 2. Furthermore, the drawing rate for the RDPO model is stable and low across different training data sizes, which shows that RDPO winning rate is higher not due to flipping the draw points but the losing points. While RDPO can increase the computation time due to the addition of rationales, the model trained on rationales can converge earlier with fewer data points than DPO. This is especially important as the cost of collecting human preference data is high. Thus, improving annotation efficiency can potentially save further training costs. Additionally, with enough computation, RDPO can reach a better model than DPO does.

Moreover, we observe for the case of the UltraFeedback dataset, with more training data, the performance of the DPO model decreases. This can be attributed to the problem of DPO overfitting and exploiting length in longer responses Azar et al. (2024); Park et al. (2024). Indeed, the UltraFeedback dataset contains chosen responses that are longer on average (1,305 character length) than the rejected responses (1,124 character length). unlike the Orca dataset (784 and 978, respectively).

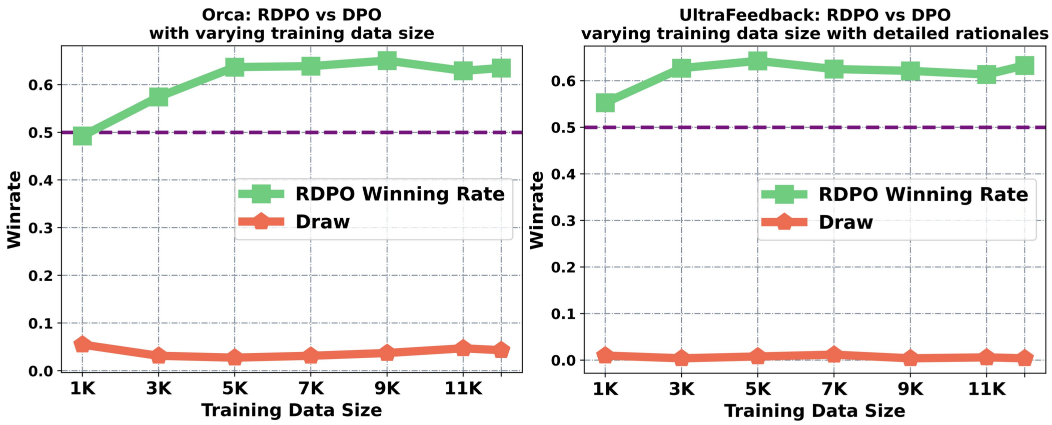

Figure 3: Winrate of RDPO model against the DPO model on respective datasets, Orca on the **left** and UltraFeedback on the **right**. The purple dashed line denotes the 0.5 mark.

Versus DPO. While we used SFT as a proxy to compare the performance between the DPO- and RDPO-trained models, here we directly compare the responses between these two models to measure the winrate for Orca and UltraFeedback datasets. We choose a DPO-trained model checkpoint for each dataset, where the winrate of the DPO model against the SFT model has converged, which is at 11,000 and 12,000, respectively. In Figure 3, we observe that the model trained with RDPO generates better responses on average than the model with DPO, even when trained with as little as 1,000 data points. With increasing training data, RDPO model improves the winrate to reach above 60% in both datasets. RDPO-trained model generates more of the preferred responses than the DPO-trained model does, even with $10\times$ fewer training points.

	Avg Output Length		TriviaQA (Exact Match)	
	DPO	RDPO	DPO	RDPO
Orca	2021	364	34.9	35.7
UltraFeedback	2066	1299	31.5	33.1

Table 1: Comparison between DPO and RDPO. **Left:** The average output lengths of the generated responses on the prompts from the test sets of respective datasets. **Right:** The exact match (EM) performance on the TriviaQA dataset of the preference-trained models on respective datasets.

Response Comparison. We compare the responses generated by the DPO- and RDPO-trained models. As shown in Table 1, the average output length by the DPO trained model is much longer than the RDPO trained model in the case of the Orca dataset, which is more than 5 times longer on average. Due to longer output, there might be a chance for a higher occurrence of hallucinations. Therefore, we want to study the correctness of the outputs from these models. For this reason, we use the TriviaQA dataset to measure the exact match (EM) accuracy and compare between the models. We see in Table 1 that models trained with the DPO loss experience a decrease in performance, compared to the models with RDPO loss. As a reason, we emphasize the importance of measuring the hallucinations in the generations of both models in future studies. We provide a comparison of the responses from the two models in Appendix B.8.

ORPO	RORPO
43	55

Table 2: Adapting rationales to the ORPO preference learning algorithm on Mistral-7B-v0.2-Instruct (Orca). Comparing the winrate of the ORPO- **(Left)** against the RORPO-trained model **(Right)**.

Adaptation to ORPO. To demonstrate the flexibility of our rationale-enriched preference learning framework, we extend the ORPO preference learning algorithm Hong et al. (2024), which omits the SFT step, to include the rationales similar to the RDPO loss, and we call it RORPO. As shown in Table 2, rationales can enhance the performance of ORPO and achieve a better winrate over the vanilla ORPO trained model. By successfully adapting rationales to both ORPO and DPO, we emphasize the simplicity of the framework as well as the effectiveness of rationales in preference learning. We further study the adaptation of these methods with rationales and evaluate the preference-trained models on the instruction-following benchmark, AlpacaEval 2.0 (Li et al., 2023b; Dubois et al., 2024), in Appendix B.4, and we observe consistent results when evaluating on this benchmark.

4.3 Rationale Quality Analysis

In this section, we examine the importance of the quality of rationales for preference learning. We study different types of rationales and possible errors encountered in rationales, and how these affect the preference learning of the model.

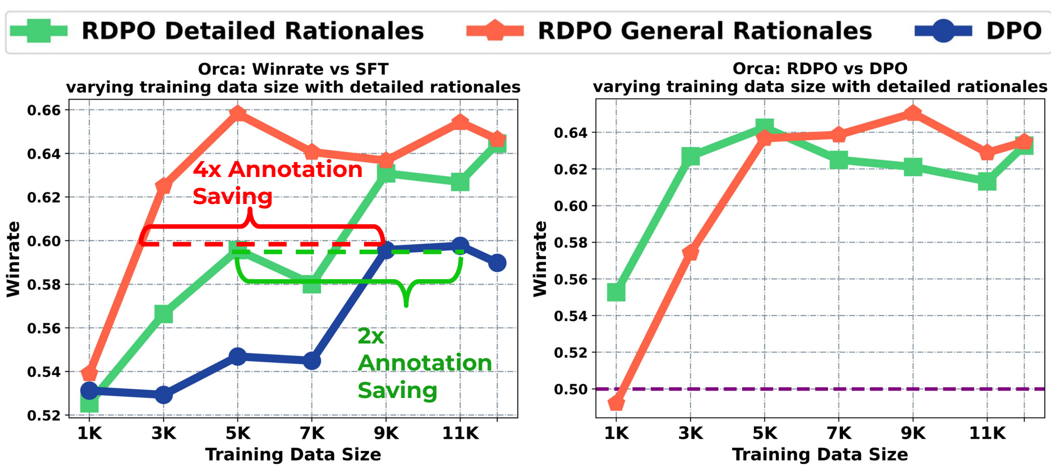

Figure 4: Measuring the impact of types of rationales on the RDPO performance. **Left:** Comparing the winrates against the SFT model. **Right:** Comparing the winrates against the DPO model.

Detailed Rationales vs General Rationales. Explaining why one answer is preferred over the other can be expressed in multiple ways through many perspectives. Here, we study the level of granularity of the rationales, general (which explains the preference at a high level without going into details) and detailed (which explains in details and pinpoints specifically to the prompt and the response). We use language models to automatically generate the rationales for the Orca dataset according to our intent. For details on generation prompts, we refer to Appendix B.5. We provide samples of these rationales in Appendix B.6. After training the models on respective rationales, we compare the winrates between RDPO trained models and the DPO one. Figure 4 on the left shows that the model trained on general rationales with the RDPO loss converges earlier to a high winning rate against the SFT model than the model trained on the detailed rationales. The reason could be that the general rationales share common features across the samples (e.g., clarity, conciseness, directness), which lets the model learn quickly and transfer these learning cues to other samples more easily, while detailed rationales might require more time to fully comprehend them. However, in both cases, the models trained on these rationales reach better winrates than the DPO against the SFT model. In

Permuted Rationales VS Original Rationales		Opposite Rationales VS Original Rationales	
<1	99	10	87

Table 3: The analysis of quality of rationales on the RDPO performance. The winrate comparison between the RDPO models trained on rationales with errors and original rationales. **Left:** Permuted, irrelevant rationales. **Right:** Opposite, inaccurate rationales.

Figure 4 (right), both RDPO models can have a better winrate $> 57\%$ against the DPO model with as few as $3{,}000$ training samples, while the DPO model is trained on $11{,}000$ samples. We provide results on models trained on additional epochs in Appendix B.3.

Low-Quality Rationales. RDPO has shown efficacy with rationales generated by the off-the-shelf models, even when the models have not undergone preference alignment. However, we want to further analyze the impact of rationale's quality on RDPO's performance. In particular, we examine the rationale quality in terms of its relevance and correctness. One case of a low-quality rationale can be a completely irrelevant rationale to the given pair of responses. To simulate irrelevant rationales, we permute the abovementioned detailed rationales over different samples so that no rationale is relevant to the context. Training the model on these rationales with RDPO and comparing one trained on original rationales, we show in Table 3 that it achieves less than 1% winrate against the RDPO model trained on correct and relevant rationales. To study the impact of correctness, we negate the general rationales to have the opposite meaning and observe that the RDPO model trained on original rationales gets almost a 90% winrate. As we note, the quality of rationales is important for improving the preference learning performance. While we showed that rationales generated by the off-the-shelf language models can already bring significant improvement to preference learning, we expect that more deliberate control of the rationale quality can further improve preference learning. We leave the in-depth exploration of strategies for generating quality-controlled rationales to future work.

GENERATED BY	RDPO VS DPO WINRATE	
MISTRAL-7B-INSTRUCT-V0.2	76 - 23	50 - 46
LLAMA3-8B-INSTRUCT	73 - 27	52 - 45
PHI3-MINI-4K	75 - 25	51 - 49
	MISTRAL-7B-INSTRUCT-V0.2	LLAMA3-8B-INSTRUCT
	TRAINED ON	

Figure 5: Studying the source of rationale generation (**Y-axis**) for the Orca dataset. Using rationales from different sources to train the models on RDPO (**X-axis**). Winrate of the RDPO against the DPO model.

Rationale Source. While collecting the human-annotated rationales could be high-quality, in practice, it is costly with time and resources. Therefore, we resolve to language models to generate rationales. In our experiments, we use the base models to create them. Here, we study the rationales coming from other sources and how they impact the RDPO training. We generate rationales for the Orca dataset on three different models: Mistral-7B-Instruct-v0.2, LLama3-8B-Instruct, and Phi3-Mini-4K Abdin et al. (2024). Then, we use these rationales to train the first two models. Results from Figure 5 show us consistent winrates against the DPO model with slightly better winrate from the same source as the base model. This shows that the rationales can be transferred to other models for preference training with rationales. Especially, leveraging models with small sizes $\{3, 7, 8\}$ billion parameters, we can generate rationales to improve preference learning.

5 A Simplified Theoretical Model for Rationales

To better understand rationale-enhanced preference learning, we employ machinaries from information theory to quantify benefits rationales provide in learning ground truth preferences under simplified assumptions. Formally, given query X, let the preference Z be a binary random variable, with $Z = 1$ indicating that response Y_1 is preferred over response Y_2, and $Z = 0$ indicating the opposite. Let R denote the rationale, $S = (X, Y_1, Y_2)$, and assume that the dataset $D = \{S_i, R_i, Z_i\}_{i=1}^n \sim \mu^n$ is sampled i.i.d. from a distribution μ.

As an intuitive first step, we quantify the benefits using the conditional mutual information between the true preferences and rationale-implied preferences given the input-response pairs, i.e., $I(Z; g(R)|S)$, where $g(R)$ capture the preference inferred from the rationale R. Intuitively, this mutual information quantity characterizes the added value of rationales in understanding true preferences, measuring how much additional information rationales provide beyond what can be inferred from input-response pairs S. Our analysis demonstrates a closed-form relationship between rationale informativeness and its alignment with true preferences (see Appendix A.1 for detailed derivations of all results).

Next, we quantify the benefits provided by rationales by directly analyzing the generalization error for preference learning with and without rationales. Compared with analyzing the informativeness of rationales outlined above, this approach offers a more direct measure of the impact of rationales on learning outcomes. Note that the generalization error of preference learning still differs from the winrate metrics used in our experiemnts for evaluating the generation of preference-aligned models. However, they are intuitively connected under the DPO loss: the minimized test loss also indicates learning the optimal data generation policy that achieves the highest reward. We defer the detailed statement of the theorem and proof to Appendix A.2. In particular, our theoretical derivation shows that training with rationales can lead to improved lower generalization error when the rationale does not contain irrelevant information other than those predictive of the preference Z, and the learning process only captures the rationale information that is useful to predict Z. Despite being derived from a simplified model, our theoretical insights align well with experimental observations. For instance, our evaluation in Section 4.3 demonstrates that detailed rationales, which may contain extraneous information, achieve lower sample efficiency compared to more general rationales that focus on preference-predictive content; furthermore, when we manually inject noise into rationales to misalign them with true preferences, we observe hampered preference learning, as indicated by lower winrates.

6 Conclusion & Limitations

In our work, we propose a paradigm shift in preference learning, emphasizing a data-centric perspective. Language models are trained by presenting them with pairs of answers, one preferred and one dispreferred, with the objective of teaching them human preferences. However, the selection of preferred answers can often be ambiguous without explanation. To address this challenge and enhance preference optimization efficiency, we introduce rationales that provide explicit reasoning for choosing one answer over another. We propose a straightforward adaptation of existing losses by incorporating these rationales into the training pipeline. Through extensive empirical experiments, we demonstrate that rationales significantly enhance training data annotation efficiency and lead to improved performance compared to baseline methods. Moreover, our approach is grounded in information theory, offering insights into how rationales enhance preference training efficiency.

To date, we have integrated rationales into our training process and successfully trained models with up to 8 billion parameters using a dataset of 12,000 samples. We encourage further investigation into the impact of rationales on preference learning, particularly exploring larger models and datasets. To facilitate research in this area, we make our code and datasets publicly available.

With the development of unpaired preference learning algorithms, such as KTO Ethayarajh et al. (2024), it is important to extend the use of rationales to handle unpaired responses in future work, e.g.,such as provided in the UltraFeedback dataset Cui et al. (2023) contains rationales for single responses without pairwise comparison.

REFERENCES

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

AI@Meta. Llama 3 model card. 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024.

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

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

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

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

Quentin Bertrand, Wojciech Marian Czarnecki, and Gauthier Gidel. On the limitations of the elo, real-world games are transitive, not additive. In *International Conference on Artificial Intelligence and Statistics*, pp. 2905–2921. PMLR, 2023.

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.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Xin Chen, Hanxian Huang, Yanjun Gao, Yi Wang, Jishen Zhao, and Ke Ding. Learning to maximize mutual information for chain-of-thought distillation. *arXiv preprint arXiv:2403.03348*, 2024.

Michelene TH Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian LaVancher. Eliciting self-explanations improves understanding. *Cognitive science*, 18(3):439–477, 1994.

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

Kevin Crowley and Robert S Siegler. Explanation and generalization in young children's strategy learning. *Child development*, 70(2):304–316, 1999.

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

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Jacob Eisenstein, Daniel Andor, Bernd Bohnet, Michael Collins, and David Mimno. Honest students from untrusted teachers: Learning an interpretable question-answering pipeline from a pretrained language model. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning*.

Kawin Ethayarajh, Winnie Xu, Dan Jurafsky, and Douwe Kiela. Human-aware loss functions (halos). Technical report, Contextual AI, 2023. https://github.com/ContextualAI/HALOs/blob/main/assets/report.pdf.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL `https://zenodo.org/records/10256836`.

Peter Hase and Mohit Bansal. When can models learn from explanations? a formal framework for understanding the roles of explanation data. *arXiv preprint arXiv:2102.02201*, 2021.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882, 2023.

Jiwoo Hong, Noah Lee, and James Thorne. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*, 2024.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1051–1068, 2023.

Shawn Im and Yixuan Li. Understanding the learning dynamics of alignment with human feedback. *arXiv preprint arXiv:2403.18742*, 2024.

Intel. Intel. `https://huggingface.co/Intel/neural-chat-7b-v3-1`, 2024. Accessed: 2024-05-18.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Yuu Jinnai and Ukyo Honda. Annotation-efficient preference optimization for language model alignment. *arXiv preprint arXiv:2405.13541*, 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2665–2679, 2023a.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`, 2023b.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*, 2022.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

Tom M Mitchell, Richard M Keller, and Smadar T Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine learning*, 1:47–80, 1986.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*, 2020.

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

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

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

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4932–4942, 2019.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pp. 1232–1240. PMLR, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*, 2024.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023.

PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. Pinto: Faithful language reasoning using prompt-generated rationales. In *The Eleventh International Conference on Learning Representations*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10266–10284, 2021.

Joseph J Williams and Tania Lombrozo. The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive science*, 34(5):776–806, 2010.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pp. 260–267, 2007.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A THEORETICAL DERIVATIONS

We begin with defining several standard quantities to be used throughout this section.

Definition 1 *Let X, Y and Z be arbitrary random variables, and let D_{KL} denote the KL divergence. We denote P_X as the marginal probability distribution of X, and $P_{Y|X}$ as the conditional distribution.*

The entropy of X is given by:

$$H(X) = -\sum_x P(X = x) \log P(X = x).$$

If X is a binary variable with $p = P(X = 1) = 1 - P(X = 0)$, then we use $H(p)$ for $H(X)$.

The joint entropy of two random variables, $H(X, Y)$, is the entropy of their joint distribution.

The conditional entropy of X given Y, $H(X|Y)$, is:

$$H(X|Y) = H(X, Y) - H(Y).$$

The mutual information between X and Y is:

$$I(X; Y) = D_{KL}(P_{X,Y} \| P_X P_Y).$$

The disintegrated mutual information between X and Y given Z is:

$$I^Z(X; Y) = D_{KL}(P_{X,Y|Z} \| P_{X|Z} P_{Y|Z}).$$

The corresponding conditional mutual information is given by:

$$I(X; Y|Z) = \mathbb{E}_Z[I^Z(X; Y)].$$

If all entropies involved are finite, it can be shown that $I(X; Y) = H(Y) - H(Y|X)$.

A.1 MUTUAL INFORMATION ANALYSIS

Formally, given query X, let the preference Z is a binary random variable, with $Z = 1$ indicating that response $Y1$ is preferred over response $Y2$, and $Z = 0$ indicating the opposite. Assume that the rationale-implied preference R is a binary random variable, with $R = 1$ indicating a rationale that supports $Y1$ being preferred, and $R = 0$ otherwise. For example, if the rationale mentions that $Y1$ is more concise and informative than $Y2$, then $R = 1$. However, there can be cases where $R \neq Z$, as the rationale may not always align perfectly with the actual preference.

For the analysis, we consider the following model: 1) The rationale R depends on the query-response (QR) pair $S = (X, Y1, Y2)$ and the preference Z, and is characterized by parameters β and α, where: $\beta = P(R = 1|Z = 1, S) = P(R = 0|Z = 0, S)$ represents the precision rate of consistency, and $\alpha = P(R = 1|Z = 0, S) = P(R = 0|Z = 1, S)$ represents the recall error due to inconsistency. 2) The preference Z is modeled as $P(Z = 1|S) = f(S) + \epsilon$, where $f(S)$ captures the preference based on the observable query-response pair S, and the additive noise term ϵ is a simple way to account for unobserved factors influencing the complex human preference that are not captured in S. *The term ϵ is referred to as "bias" in the* ***main text*** *that accounts for the difference between the true preference and prediction based on the query and responses alone.*

Theorem 1 *Under the given assumptions, the mutual information $I(Z; R|S)$ is given by:*

$$H(p + \epsilon) - (\beta(p + \epsilon) + \alpha(1 - p - \epsilon)) \cdot H\left(\frac{\beta(p + \epsilon)}{\beta(p + \epsilon) + \alpha(1 - p - \epsilon)}\right)$$
$$- (1 - (\beta(p + \epsilon) + \alpha(1 - p - \epsilon))) \cdot H\left(\frac{\alpha(p + \epsilon)}{\alpha(p + \epsilon) + \beta(1 - p - \epsilon)}\right),$$

where $p = f(S)$. The mutual information $I(Z; R|S)$ satisfies the following properties in three distinct regimes:

1. Uninformative rationale regime: If $\beta = \alpha = 0.5$, then $I(Z; R|S) = 0$.

2. Maximally informative rationale regime: If $\beta = 1$ and $\alpha = 0$, then $I(Z; R|S) = H(p + \epsilon)$.

3. Moderately informative rationale regime: If $\beta = 0.5 + \gamma$ and $\alpha = 0.5 - \gamma$, where $0 < \gamma < 0.5$, then $I(Z; R|S)$ increases with γ, ranging from 0 when $\gamma = 0$ (uninformative rationale) to $H(p + \epsilon)$ when $\gamma = 0.5$ (maximally informative rationale).

The theorem highlights that the potential benefits of including rationales in the training process for preference learning tasks may vary in different regimes.

Regime 1: Highly informative rationale ($\beta = 1$ and $\alpha = 0$): In this case, the mutual information is solely determined by the entropy of the preference prediction from the query and responses, $H(f(S) + \epsilon)$. Let us interpret $f(S)$ as the query-response-dependent (QR-dependent) confidence generator that only depends on S, and ϵ captures the idiosyncrasies such as unknown confounding factors that influence the probability of preference. If the true probability $P(Z = 1|S) = f(S) + \epsilon$ is less than 0.5, i.e., Z is most likely to be 0, a positive $\epsilon > 0$ means that the true probability $P(Z = 1|S)$ is less extreme than the confidence score $f(S)$ as it gets closer to 0.5 with increasing ϵ. On the contrary, a negative $\epsilon < 0$ means that the true probability $P(Z = 1|S)$ is more extreme than the QR-dependent confidence $f(S)$ as it gets closer to 0 with increasing $|\epsilon|$.

From the perspective of the QR-dependent confidence generator $f(S)$ that tries to explain the preference based on QR pairs, a positive $\epsilon > 0$ would make it look more confident than it should be, i.e., overconfident, while a negative $\epsilon < 0$ would make it less confident (or more conservative) based on the QR sequence than it should be.

If it is likely to be overconfident based on the QR pairs, i.e., $\epsilon > 0$, then the more positive ϵ is, the more risk there is of being overconfident in the QR-dependent predictor $f(S)$. In this case, there is a lot of mutual information in $I(Z; R|S)$, so having rationales can "soften" the potential overconfidence by bringing additional information other than the QR pair, which would otherwise occur in QR pairs alone, as in traditional reward modeling. Similar analysis holds when $P(Z = 1|S) = f(S) + \epsilon$ is greater than 0.5, i.e., Z is most likely to be 1.

Key message: Rationales are most useful when the reward modeling based on QR alone tends to have bias (i.e., overconfident).

Regime 2: Uninformative rationale ($\beta = \alpha = 0.5$): In this regime, the rationale provides no *additional* information about the preference, and the mutual information $I(Z; R|S)$ is zero.

Regime 3: Moderately informative rationale (high precision $\beta = 0.5 + \gamma$ and low recall error $\alpha = 0.5 - \gamma$, where $0 < \gamma < 0.5$): In this regime, it can be shown based on dereivative analysis that as γ increases (more informative rationale), the terms involving γ in the numerators and denominators of the conditional entropies become more prominent. The mutual information will increase with γ, as the rationale becomes more informative about the preference.

A.2 Theorem 2: Generalization Bound

Next, we analyze the sample complexity of training language models with and without rationales to predict preferences. We consider two regimes: *1)* **Training with rationale**: Let $\theta_{\text{ra}} = \mathcal{A}_{\text{ra}}(D) \sim P_{\theta_{\text{ra}}|D}$ denote the parameters of the language model trained to predict Z given S and R. *2)* **Training without rationale**: Let $\theta_{\text{un}} = \mathcal{A}_{\text{un}}(D_{\setminus R}) \sim P_{\theta_{\text{un}}|D_{\setminus R}}$ denote the output parameters trained to predict Z given only S, where $D_{\setminus R}$ is a dataset D with rationales removed. Given a loss function ℓ that measures the prediction of preference Z, the (mean) generalization error is $\text{gen}(\mu, \mathcal{A}) = \mathbb{E}_{D,\theta\sim\mathcal{A}(D)}|\mathbb{E}_{\mu}[\ell(\theta)] - \mathbb{E}_D[\ell(\theta)]|$, where $\mathbb{E}_{\mu}[\ell(\theta)]$ is the expected loss on the true distribution (true risk) and $\mathbb{E}_D[\ell(\theta)]$ is the empirical risk.

We introduce the following conditions on the relationship between S, R, and Z, and the learning process: *1)* $H(R|Z) \leq \eta_1$ and $H(Z|R) \leq \eta_2$, i.e., the rationale R is informative about Z (small η_2) without excessive irrelevance (small η_1). *2)* $I(\theta_{\text{ra}}; S|Z, R) \leq \delta$ for some small positive constant δ, i.e., the learned model θ_{ra} does not capture much additional information from S beyond what Z and R already provide. Condition 1 is supported by an effective procedure to generate useful rationale R. To justify Condition 2, if the learning algorithm is designed to focus on capturing the information in R, which is highly informative about Z (per Condition 1 for small η_2), we can show that the model

θ_{ra} can accurately predict Z without *needing* to capture much additional information from S beyond what is already present in Z and R. We provide rigorous but partial justification in Appendix A.3.

Theorem 2 (Generalization bounds) *Suppose the loss function ℓ is σ-subgaussian under the true data distribution. Under conditions 1 and 2, we have:*

$$\operatorname{gen}(\mu, \mathcal{A}_{ra}) \leq \sqrt{\frac{2\sigma^2}{n} \cdot (I(\theta_{ra}; Z) + \delta + \eta_1)}, \tag{5}$$

$$\operatorname{gen}(\mu, \mathcal{A}_{un}) \leq \sqrt{\frac{2\sigma^2}{n} \cdot (I(\theta_{un}; Z) + I(\theta_{un}; S|Z))}. \tag{6}$$

The proof relies on the mutual information-based generalization bounds (Russo & Zou, 2016; Xu & Raginsky, 2017) and the decomposition of the mutual information terms for both training regimes using the chain rule (see Appendix A.2). The terms $I(\theta_{\text{ra}}; Z)$ and $I(\theta_{\text{un}}; Z)$ can be expected to be similar as long as both regimes achieve good prediction of Z. Under the conditions of the theorem, we can observe that the sample complexity reduction depends on the gap between $I(\theta_{\text{un}}; S|Z)$ and $\delta + \eta_1$; training with rationales can lead to improved sample efficiency when the rationale does not contain irrelevant information other than those predictive of the preference Z, i.e., η_1 is small, and the learning process only captures the rationale information that is useful to predict Z. The theoretical insights are supported by our experimental results. For instance, our evaluation in Section 4.3 demonstrates that a detailed rationale achieves lower sample efficiency compared to more general rationales, which contain less irrelevant information beyond what is predictive of the preference; furthermore, we showed that irrelevant rationales, i.e., a large value of η_1, indeed hamper learning.

For the proof, recall that given a loss function ℓ, the (mean) generalization error is $\operatorname{gen}(\mu, \mathcal{A}) = \mathbb{E}_{D,\theta\sim\mathcal{A}(D)}|\mathbb{E}_\mu[\ell(\theta)] - \mathbb{E}_D[\ell(\theta)]|$, where $\mathbb{E}_\mu[\ell(\theta)]$ is the expected loss on the true distribution (true risk) and $\mathbb{E}_D[\ell(\theta)]$ is the empirical risk. For fair comparison between $\operatorname{gen}(\mu, \mathcal{A}_{\text{un}})$ and $\operatorname{gen}(\mu, \mathcal{A})_{\text{ra}}$, some technical nuances arise. The key difference from the typical setup in Xu & Raginsky (2017) is that the true data distribution μ includes the distribution for the rationale R, but the training regime $\operatorname{gen}(\mu, \mathcal{A}_{\text{un}})$ does not explictly use this information. However, it may seem unclear initially whether we should include that in the generalization bound, since R is indeed generated based on Z, corresponding to the true Markov chain: $S \to Z \to R$.

To clarify, the Markov chain for the training without rationale is: $S \to Z \to R$, with additional arrows $Z \to \theta_{\text{un}}$ and $S \to \theta_{\text{un}}$, but no arrow from R to θ_{un}.

Intuitively, we should account for this difference by arguing that, conditioned on the preference Z, the learned model θ_{un} is conditionally independent of R. However, due to this difference, it seems prudent to reason from first principles.

Let's start by choosing the distributions P and Q for the Donsker-Varadhan variational representation of the KL divergence. We set $P = P_{S,R,Z,\theta_{\text{un}}}$ and $Q = \mu^n \otimes P_{\theta_{\text{un}}}$, where μ is the distribution for (S, R, Z). Then, for any measurable function f, we have:

$$D(P\|Q) \geq \mathbb{E}_P[f(S, R, Z, \theta)] - \log \mathbb{E}_{(\bar{D},\bar{R},\bar{Z},\bar{\theta})\sim Q}[e^{f(\bar{S},\bar{R},\bar{Z},\bar{\theta})}]. \tag{7}$$

Now, choose $f(S, R, Z, \theta) = \lambda(\ell_D(\theta) - \ell_\mu(\theta))$ for some $\lambda \in \mathbb{R}$, where $\ell_D(\theta)$ is the empirical loss on the dataset D and $\ell_\mu(\theta)$ is the expected loss under the true distribution μ. Substituting this into equation 7, we get:

$$\begin{aligned} D(P\|Q) &\geq \lambda(\mathbb{E}[\ell_D(\theta)] - \mathbb{E}[\ell_\mu(\theta)]) - \log \mathbb{E}_{(\bar{D},\bar{R},\bar{Z},\bar{\theta})\sim Q}[e^{\lambda(\ell_{\bar{D}}(\bar{\theta}) - \ell_\mu(\bar{\theta}))}] \\ &\geq \lambda(\mathbb{E}[\ell_D(\theta)] - \mathbb{E}[\ell_\mu(\theta)]) - \frac{\lambda^2\sigma^2}{2n}, \end{aligned} \tag{8}$$

where the second inequality follows from the fact that $\ell_{\bar{D}}(\bar{\theta})$ is $\sigma/\sqrt{n}$-subgaussian under Q, due to the subgaussian assumption on the loss function.

As equation 8 holds for any $\lambda \in \mathbb{R}$, it must also hold for the λ that minimizes the right-hand side. This minimum occurs at $\lambda^* = n(\mathbb{E}[\ell_D(\theta)] - \mathbb{E}[\ell_\mu(\theta)])/\sigma^2$, yielding:

$$D(P\|Q) \geq \frac{n}{2\sigma^2}(\mathbb{E}[\ell_D(\theta)] - \mathbb{E}[\ell_\mu(\theta)])^2.$$

Taking the square root of both sides, we have:

$$\mathrm{gen}(\mu, \mathcal{A}_{\text{un}}) \le \sqrt{\frac{2\sigma^2}{n} D(P\|Q)}.$$

The key observation here is that $P_{S,R,Z,\theta_{\text{un}}} = P_{\theta_{\text{un}}|S,R,Z} P_{S,R,Z} = P_{\theta_{\text{un}}|S,Z} P_{S,R,Z}$, since θ_{un} is conditionally independent of R given S and Z in this training regime. Therefore,

$$\begin{aligned} D_{\text{KL}}(P\|Q) &= D_{\text{KL}}(P_{S,R,Z,\theta_{\text{un}}} \| \mu^n \otimes P_{\theta_{\text{un}}}) \\ &= D_{\text{KL}}(P_{\theta_{\text{un}}|S,Z} P_{S,Z} \| \mu^n_{\setminus R} \otimes P_{\theta_{\text{un}}}) \\ &= I(S, Z; \theta_{\text{un}}), \end{aligned}$$

where $\mu_{\setminus R}$ denotes the marginal distribution of μ over S and Z (i.e., excluding R).

Hence, we have:

$$\mathrm{gen}(\mu, \mathcal{A}_{\text{un}}) \le \sqrt{\frac{2\sigma^2}{n} \cdot I(S, Z; \theta_{\text{un}})}.$$

Note that we use $I(S, Z; \theta_{\text{un}})$ instead of $I(S, R, Z; \theta_{\text{un}})$, which is the key difference from the typical setup in Xu & Raginsky (2017), where the learned model is assumed to depend on all the data.

We can then arrive at the result for training without rationale by noting:

$$I(\theta_{\text{un}}; S, Z) = I(\theta_{\text{un}}; Z) + I(\theta_{\text{un}}; S|Z).$$

For training with rationale, we have:

$$\begin{aligned} I(\theta_{\text{ra}}; S, R, Z) &= I(\theta_{\text{ra}}; Z) + I(\theta_{\text{ra}}; R|Z) + I(\theta_{\text{ra}}; S|Z, R) \\ &\le I(\theta_{\text{ra}}; Z) + H(R|Z) + I(\theta_{\text{ra}}; S|Z, R) \\ &\le I(\theta_{\text{ra}}; Z) + \eta_1 + \delta, \end{aligned}$$

where the first inequality is due to $I(\theta_{\text{ra}}; R|Z) \le H(R|Z)$, and the second inequality follows from conditions 1 and 2. We can now apply (Xu & Raginsky, 2017, Thm. 1) to yield the result.

A.3 Supporting Lemmas

Lemma 1 *Let $\hat{Z}$ be the estimate of Z based on θ. Let $P_e = P(Z \ne \hat{Z}|\hat{Z}, \theta)$ be the probability of error in predicting Z using θ. Suppose $P_e \le \epsilon$. Then, we have that:*

$$P_e \ge \frac{H(Z) - I(R;\theta) - I(Z;\theta|R) - H(\epsilon)}{\log|Z|}. \tag{9}$$

Proof: First, let's define an indicator variable E for the error event: $E = \begin{cases} 0, & \text{if } \hat{Z} = Z \\ 1, & \text{if } \hat{Z} \ne Z \end{cases}$. We have

$$\begin{aligned} H(Z|\theta) = H(Z|\hat{Z}, \theta) &= H(Z, E|\hat{Z}, \theta) \\ &= H(E|\hat{Z}, \theta) + H(Z|E, \hat{Z}, \theta) \\ &= H(P_e) + P_e H(Z|E = 1, \hat{Z}, \theta) + (1 - P_e) H(Z|E = 0, \hat{Z}, \theta) \\ &\le H(\epsilon) + P_e \log|Z|. \end{aligned}$$

Hence, we have

$$P_e \ge \frac{H(Z|\theta) - H(\epsilon)}{\log|Z|} = \frac{H(Z) - I(Z;\theta) - H(\epsilon)}{\log|Z|}. \tag{10}$$

This part of the proof follows from Fano's inequality.

Now, since $I(Z;\theta) + I(R;\theta|Z) = I(Z, R;\theta)$, we have

$$\begin{aligned} I(Z;\theta) &= I(Z, R;\theta) - I(R;\theta|Z) \\ &= I(R;\theta) + I(Z;\theta|R) - I(R;\theta|Z) \\ &\le I(R;\theta) + I(Z;\theta|R) \end{aligned}$$

Plugging in equation 10 obtains the result.

Note: $I(Z;\theta|R) = 0$ if θ is trained only on R, i.e., Z and θ are conditionally independent given R.

Lemma 2 *Let $\hat{Z}$ be the estimate of Z based on θ. Let $P_e = P(Z \neq \hat{Z}|\hat{Z}, \theta)$ be the probability of error in predicting Z using θ. Assume that $P_e \leq 0.5$. Then, we have that:*

$$P_e \leq \frac{H(Z) - I(R;\theta) + H(R|Z)}{\log 2}. \tag{11}$$

Proof: By definition of E from Lemma 1, we have

$$\begin{aligned} H(Z|\theta) &= H(Z|\hat{Z},\theta) = H(Z,E|\hat{Z},\theta) \\ &= H(E|\hat{Z},\theta) + H(Z|E,\hat{Z},\theta) \\ &\geq H(E|\hat{Z},\theta) \\ &= P_e \log(1/P_e) + (1-P_e)\log(1/(1-P_e)) \\ &\geq P_e \log 2 \end{aligned}$$

Hence, we have $P_e \leq H(Z|\theta)/\log 2$. Since $H(Z|\theta) = H(Z) - I(Z;\theta)$, and that

$$\begin{aligned} I(Z;\theta) &= I(Z,R;\theta) - I(R;\theta|Z) \\ &= I(R;\theta) + I(Z;\theta|R) - I(R;\theta|Z) \\ &\geq I(R;\theta) - H(R|Z) + H(R|\theta,Z) \\ &\geq I(R;\theta) - H(R|Z), \end{aligned}$$

the bound follows.

Partial justification of Condition 2: Consider the Markov chain $Z \to R \to \theta$ with additional arrows of $Z \to \theta$ and $R \to \theta$, where Z is the preference, R is the rationale, and θ is a trained model used to predict Z. From both Lemma 1 and 2, we see that as long as θ captures the information of R, i.e., $I(R;\theta)$ is large, and R does not contain excessive irrelevant information other than Z, i.e., $H(R|Z)$ is small, the prediction error of Z from model θ can be well-controlled. Specifically, based on the lower and upper bound analysis, we can conclude that the probability of error P_e decreases with increasing $I(R;\theta)$ or decreasing $H(R|Z)$. This implies that the model does not need to capture additional information from S to achieve high prediction accuracy for Z, i.e., $I(\theta_{\text{ra}}; S|Z,R)$ is small. In other words, the incorporation of R in the training of θ_{ra} guides the model to easily predict Z without resorting to finding potentially irrelevant information from S. A full justification of the condition hinges on a detailed analysis of the specific algorithm and is beyond the scope of this study.

A.4 Proof of Theorem 1 and Derivation of Regimes

Recall the definition of $\alpha, \beta, \epsilon, f(\cdot)$ in Sec. A.1. Now, we derive the relationship between the mutual information $I(Z;R|S)$ and these parameters:

$$\begin{aligned} P(Z=1|S,R=1) &= \frac{P(R=1|Z=1,S)P(Z=1|S)}{P(R=1|S)} \\ &= \frac{P(R=1|Z=1,S)P(Z=1|S)}{\sum_{z\in\{0,1\}} P(R=1|Z=z,S)P(Z=z|S)} \\ &= \frac{\beta(f(S)+\epsilon)}{\beta(f(S)+\epsilon) + \alpha(1-f(S)-\epsilon)}. \end{aligned}$$

Similarly, we have

$$P(Z=1|S,R=0) = \frac{\alpha(f(S)+\epsilon)}{\alpha(f(S)+\epsilon) + \beta(1-f(S)-\epsilon)}.$$

These equations show that the probability of $Z=1$ given the query X, the preferred response Y_1, the dispreferred response Y_2, and the rationale R depends on both the informativeness of the rationale, through α and β, and the informativeness of the query and responses, through $f(S)$. Using the above equations, we get the conditional entropies as follows:

$$\begin{aligned} H(Z|S,R=1) &= H\left(\frac{\beta(f(S)+\epsilon)}{\beta(f(S)+\epsilon) + \alpha(1-f(S)-\epsilon)}\right), \\ H(Z|S,R=0) &= H\left(\frac{\alpha(f(S)+\epsilon)}{\alpha(f(S)+\epsilon) + \beta(1-f(S)-\epsilon)}\right) \end{aligned}$$

and substituting to the mutual information equation, we get:

$$I(Z;R|S) = H(Z|S) - \sum_{r=\{0,1\}} P(R=r|S)H(Z|S,R=r).$$

Then, we compute each probabilities $P(R|S)$ as follows:

$$\begin{aligned}
P(R=1|S) &= \sum_{z\in\{0,1\}} P(R=1|Z=z,S)P(Z=z|S) \\
&= \beta(f(S)+\epsilon) + \alpha(1-f(S)-\epsilon), \\
P(R=0|S) &= 1 - P(R=1|S) \\
&= 1 - (\beta(f(S)+\epsilon) + \alpha(1-f(S)-\epsilon))
\end{aligned}$$

and substitute them back into the conditional mutual information term, where we define $p = f(S)$:

$$\begin{aligned}
I(Z;R|S) = H(p+\epsilon) - (\beta(p+\epsilon) + \alpha(1-p-\epsilon))H\left(\frac{\beta(p+\epsilon)}{\beta(p+\epsilon)+\alpha(1-p-\epsilon)}\right) \\
- (1-(\beta(p+\epsilon)+\alpha(1-p-\epsilon)))H\left(\frac{\alpha(p+\epsilon)}{\alpha(p+\epsilon)+\beta(1-p-\epsilon)}\right).
\end{aligned}$$

To study the influence of the parameters α, β, ϵ on the mutual information, we consider the following edge cases:

Regime 1: Highly informative rationale $\beta \approx 1$ and low noise $\alpha \approx 0 \implies$ Rationale is a sufficient statistics.
In this case, the conditional probabilities are simplified to

$$\begin{aligned}
P(Z=1|S,R=1) &\approx \frac{f(S)+\epsilon}{f(S)+\epsilon} = 1, \\
P(Z=1|S,R=0) &\approx \frac{0}{1-f(S)-\epsilon} = 0.
\end{aligned}$$

Thus, the mutual information becomes as follows:

$$\begin{aligned}
I(Z;R|S) \approx H(f(S)+\epsilon) - P(R=1|S)H(1) \\
-P(R=0|S)H(0) = H(f(S)+\epsilon).
\end{aligned}$$

In this regime, the conditional mutual information is solely determined by the entropy of the preference prediction, $H(f(S)+\epsilon)$. We notice that the entropy function is concave and reaches the max value at the 0.5 mark.

Regime 2: Uninformative rationale $\beta \approx 0.5$ and high noise $\alpha \approx 0.5$.

For this case, the conditional probabilities become:

$$P(Z=1|S,R=1) \approx \frac{f(S)+\epsilon}{f(S)+\epsilon+1-f(S)-\epsilon} = f(S)+\epsilon = P(Z=1|S,R=0)$$

and the mutual information equals to:

$$I(Z;R|S) \approx H(f(S)+\epsilon) - H(f(S)+\epsilon) = 0,$$

which shows that the rationales provides 0 information about the preference given the prompt and responses.

Regime 3: Moderately informative rationale $\beta = 0.5+\gamma$ and lower noise $\alpha = 0.5-\gamma$.

Given the assumption of $\beta = 0.5+\gamma$ and $\alpha = 0.5-\gamma$, where $0 \le \gamma \le 0.5$ and γ denotes the level of informativeness of the rationale, we substitute this into the conditional mutual information term.

We first substitute into the conditional probabilities and get:

$$P(Z=1|S,R=1) = \frac{(0.5+\gamma)(f(S)+\epsilon)}{(0.5+\gamma)(f(S)+\epsilon)+(0.5-\gamma)(1-f(S)-\epsilon)},$$
$$P(Z=1|S,R=0) = \frac{(0.5-\gamma)(f(S)+\epsilon)}{(0.5-\gamma)(f(S)+\epsilon)+(0.5+\gamma)(1-f(S)-\epsilon)}.$$

Then, we compute the following probabilities:

$$\begin{aligned} P(R=1|S) &= \sum_{z\in\{0,1\}} P(R=1|Z=z,S)P(Z=z|S) \\ &= 0.5+2\gamma(f(S)+\epsilon-0.5), \\ P(R=0|S) &= 0.5-2\gamma(f(S)+\epsilon-0.5). \end{aligned}$$

Now, we can compute the conditional mutual information term:

$$\begin{aligned} I(Z;R|S) = H(f(S)+\epsilon) \\ &- (0.5+2\gamma(f(S)+\epsilon-0.5))\cdot H\left(\frac{(0.5+\gamma)(f(S)+\epsilon)}{(0.5+\gamma)(f(S)+\epsilon)+(0.5-\gamma)(1-f(S)-\epsilon)}\right) \\ &- (0.5-2\gamma(f(S)+\epsilon-0.5))\cdot H\left(\frac{(0.5-\gamma)(f(S)+\epsilon)}{(0.5-\gamma)(f(S)+\epsilon)+(0.5+\gamma)(1-f(S)-\epsilon)}\right). \end{aligned} \tag{12}$$

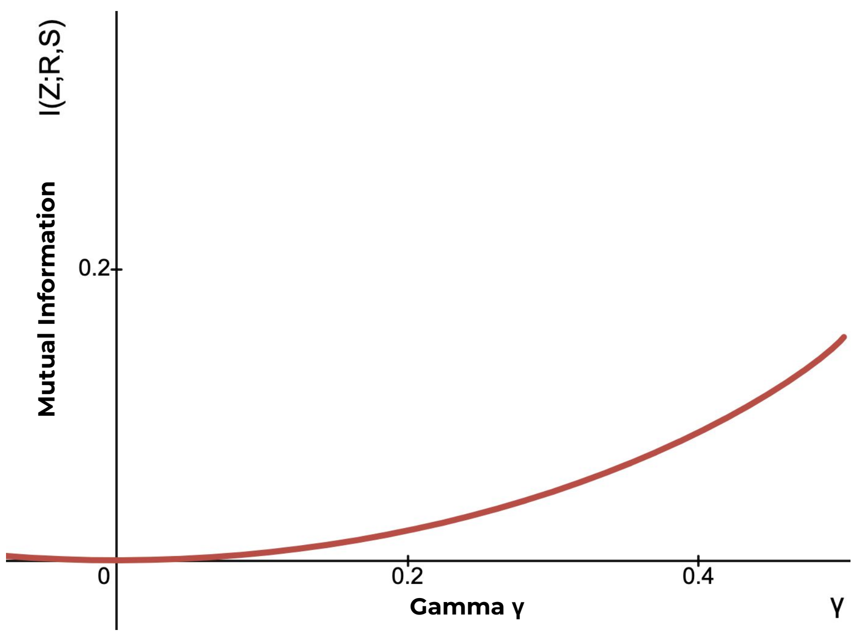

Figure 6: The plot of Equation 12 showing the relation between mutual information and gamma γ for a fixed $f(S)$.

We can now analyze the behavior of the mutual information as a function of γ:

When the rationale is uninformative $\gamma=0$, then the mutual information becomes 0, $I(Z;R|S)=0$, which is consistent with previous cases, in which uninformative rationales provide no additional information about the preference Z as demonstrated in Figure 6.

As rationale becomes more informative about the preference by increasing γ, we observe that mutual information also increases displayed in Figure 6.

Consider the case that the true probability $P(Z=1|S)=f(S)+\epsilon>0.5$, so the preference Z is most likely to be 1.

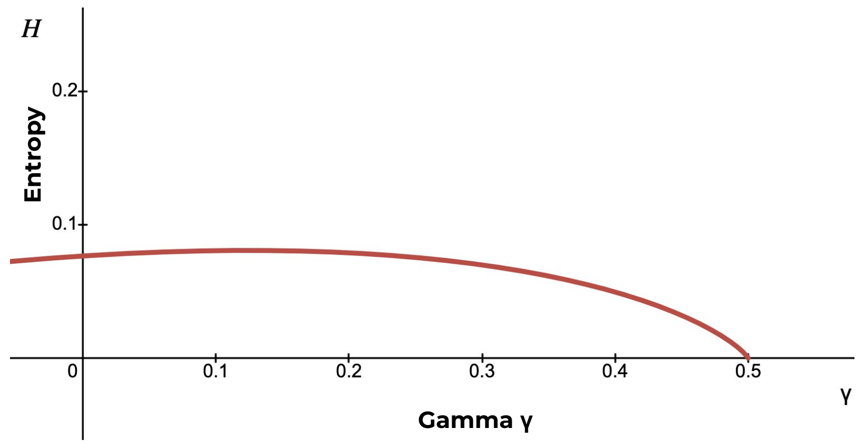

Figure 7: The plot of Equation 12 showing the relation between the first entropy term and gamma γ for a fixed $f(S)$.

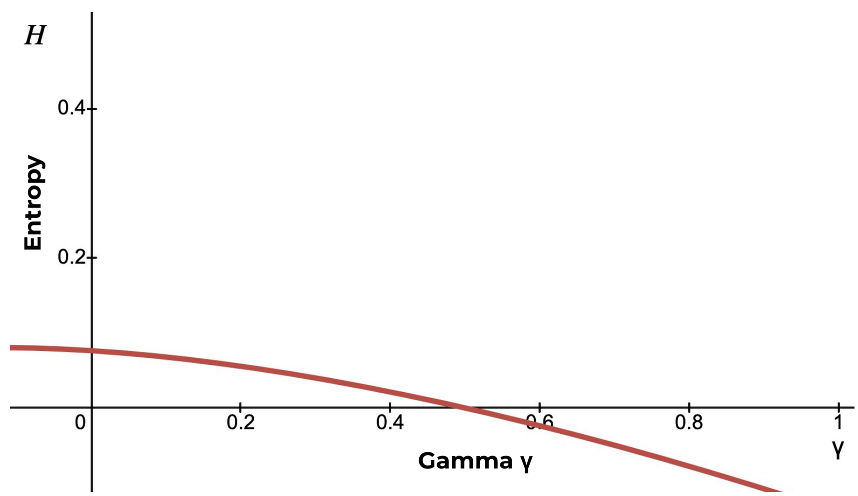

Figure 8: The plot of Equation 12 showing the relation between second entropy term and gamma γ for a fixed $f(S)$.

Then, we now focus on terms in $I(Z;R|S)$ that contain γ. As γ increases, the first entropy weight term $(0.5 + 2\gamma(f(S) + \epsilon - 0.5))$ increases and the second entropy weight term $(0.5 - 2\gamma(f(S) + \epsilon - 0.5))$ decreases. Entropy terms also involve γ, and we observe that as γ increase, the ratio $\frac{(0.5+\gamma)(f(S)+\epsilon)}{(0.5+\gamma)(f(S)+\epsilon)+(0.5-\gamma)(1-f(S)-\epsilon)}$ approaches 1, since the numerator grows faster than the denominator. Thus, the first entropy term decreases with γ, as the entropy of a distribution to a deterministic one is lower.

Conversely, for the second entropy term, with the increase of γ, the ratio $\frac{(0.5-\gamma)(f(S)+\epsilon)}{(0.5-\gamma)(f(S)+\epsilon)+(0.5+\gamma)(1-f(S)-\epsilon)}$ approaches 0, so the second entropy term also decreases with γ.

- For the first entropy term, with an increase of γ, the weight of the first entropy term increases, but the entropy decreases itself (see Figure 7).
- For the second entropy term, with an increase of γ, the weight of the second entropy term decreases, and the entropy decreases itself (see Figure 8).

The net effect on the mutual information depends on the relative magnitudes of these changes. However, we can argue that the decrease in the entropy terms dominates the change in their weights due to the entropy function changing more rapidly near the extremes (i.e., when the distribution is close to being deterministic) compared to the middle range.

Thus, with an increase in γ, the overall contribution of the entropy terms to the mutual information decreases, causing an increase in $I(Z;R|S)$, which indicates that as the rationale becomes more informative about the preference, the mutual information increases.

B Additional Experimental Results

B.1 Experimental Details

For DPO-based methods, we fine-tune the base model by supervised fine-tuning (SFT) with the chosen responses from the preference dataset for a single epoch. For ORPO, which avoids the reference model, we skip this SFT step. For models trained on RDPO and results reported in Section 4, we use $\gamma = 2.0$, and for RORPO, we use $\gamma = 10.0$. Similar to baseline methods, we train RDPO and RORPO for 1 epoch. We perform ablation studies on the hyperparameter γ and the number of epochs in the following sections.

For our winrate scores, we report the mean winrates after querying the evaluator 3 times. To reduce the evaluator's order bias, we have additionally shuffled the order of responses. We note that the winrate error bars are within $< 3\%$ for 512 samples.

B.2 Ablation Study on Models and Hyperparameters

	Mistral-7B-v0.1					
General	62	61	63	61	62	60
Detailed	64	66	63	61	60	62
γ	1.0	1.5	2.0	2.5	3.0	10.0

	Mistral-7B-v0.2-Instruct					
General	55	62	57	55	57	56
Detailed	56	56	57	60	57	59
γ	1.0	1.5	2.0	2.5	3.0	10.0

	Zephyr-7B-Beta					
General	58	55	55	57	55	57
Detailed	48	49	51	51	50	51
γ	1.0	1.5	2.0	2.5	3.0	10.0

Table 4: The impact of different values of hyperparameter γ on the winrate of the RDPO model against the DPO model. The results on various models: Mistral-7B-v0.1 **(Top)**, Mistral-7B-v0.2-Instruct **(Middle)**, and Zephyr-7B-Beta **(Bottom)**.

Here, we investigate the impact of the hyperparameter γ, ranging from 1.0 to 10.0, on the performance of the model trained with RDPO loss. We provide the winrate scores against the DPO model on the Orca dataset. As we see in Table 4, models trained on either general or detailed rationales can still achieve a stronger winrate against the DPO model, except for the case of *Zephyr-7B-Beta* model, which achieves a draw with the DPO model. For this model, the better quality rationales are important for effective preference learning, as the general rationales can still improve the performance.

B.3 Ablation on the Number of Epochs

In the main paper, we trained the models with the RDPO loss on a single epoch similar to DPO (Rafailov et al., 2024). Here, we study the impact of training the models with the rationales on more epochs. As we observe in Table 5, training with more epochs does not improve the

	General	Detailed
Epoch 1	76	74
Epoch 2	75	71
Epoch 3	74	74

Table 5: The analysis of the number of epochs on the performance. Winrate of the RDPO model trained on the general rationales **(left)** and detailed rationales **(right)** against the DPO model, respectively.

winrate of the RDPO model against the DPO model. The reason could be that the model has already learned the preference well after the first epoch, or it could be that the quality of rationales could be further improved to increase the efficiency of the rationale-based preference learning algorithms.

B.4 Evaluation of the Impact of Rationales with AlpacaEval 2.0

Here, we study the flexibility of our method by extending different preference learning methods, such as DPO (Rafailov et al., 2024) and ORPO (Hong et al., 2024), with rationales. In this experiment, we evaluate the performance of the trained models on the automatic instruction-following AlpacaEval 2.0 benchmark (Li et al., 2023b) with GPT-4-turbo as a judge and report the raw winrate, the length-controlled (LC) winrate (Dubois et al., 2024), which is robust against the verbosity bias that the raw winrate inherently entails, and the average response length. We train the instruction-tuned models, Mistral-7B-Instruct-v0.2 and Llama-3-8B-Instruct, on the Intel-DPO-Pairs dataset.

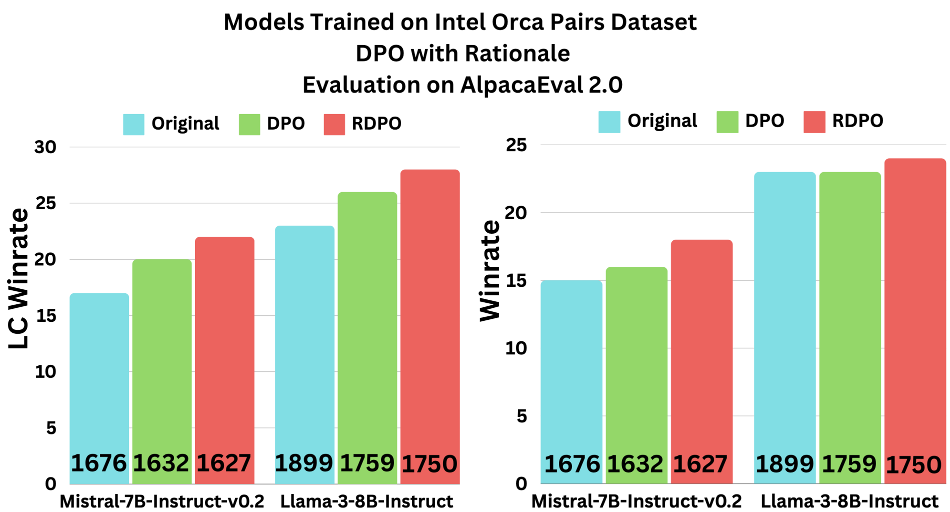

Figure 9: The performance comparison of the original model, DPO trained model, and DPO with rationales (RDPO) trained model on the AlpacaEval 2.0 benchmark. The **bolded** numbers denote the average response length of each models.

We observe in Figure 9 that DPO trained model improves the winrates on both models compared to the original model. Additionally, RDPO models further increase the win rates. We note in Figure 10 that in the case of the Mistral model, the winrate decrease with ORPO preference training, which might be due to the lack of the reliance on the reference model, the behavior which is also observed in Ethayarajh et al.. Furthermore, after adding rationales, the (LC) winrate not only increases but also surpasses of the original model. These results show the helpfulness of adding rationales into preference learning.

Interestingly, we also note that rationale based models increase the winrates while their average response lengths are decreased compared to average response lengths of the original model, which is not a similar observation as seen in some current methods, such as SPPO (Wu et al., 2024) or SimPO (Meng et al., 2024).

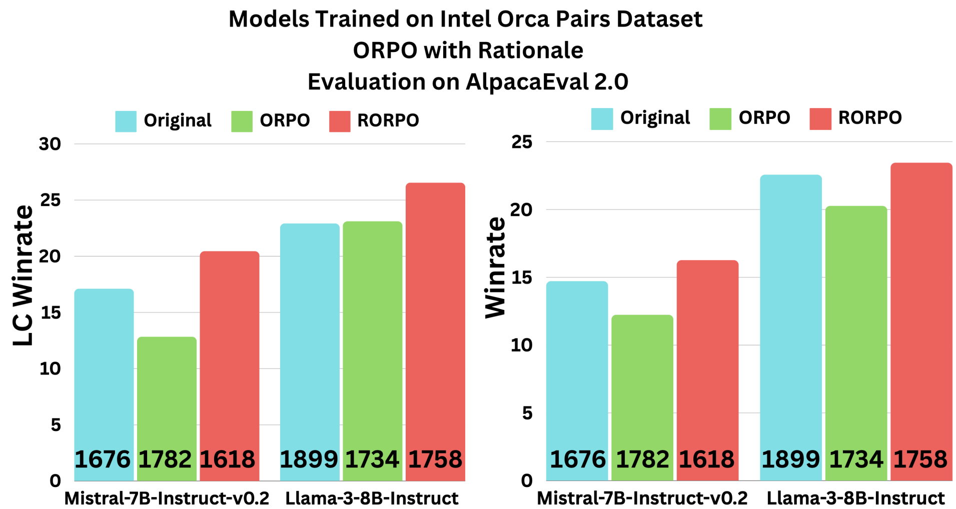

Figure 10: The performance comparison of the original model, ORPO trained model, and ORPO with rationales (RORPO) trained model on the AlpacaEval 2.0 benchmark. The **bolded** numbers denote the average response length of each models.

B.5 Rationale Generation

Here, we provide the prompts to generate detailed and general rationales as mentioned in Section 4. In Prompt 1, we ask the model to generate general rationales without going into details and provide high-level reasoning. In Prompt 2, we ask the model to provide a specific rationale and go as much into detail as possible. These are our examples of prompts and we call for further studies to improve the prompts for higher quality rationales.

```
Given the prompt:
<prompt>
Which of the following assistant's responses is preferred and
strictly follows the prompt question?  Why this assistant's
response:
<chosen>
is preferred over the response provided below:
<rejected>
On a high and general level, why the response above is
preferred over the response below?  Provide a general,
high-level explanation for your reasoning without going into
the response's details.
```

Prompt 1: Prompt to the LLM to retrieve a general rationale.

```
Given the prompt:
<prompt>
Which of the following assistant's responses is preferred and
strictly follows the prompt question?  Why this assistant's
response:
<chosen>
is preferred over the response provided below:
<rejected>
Why the response above is preferred over the response below?
Provide a specific, detailed explanation for your reasoning
pointing directly at the response.
```

Prompt 2: Prompt to the LLM to retrieve a detailed rationale.

B.6 COMPARISON OF RATIONALES

We provide examples of the general and detailed rationales for the Orca dataset by the Mistral-7B-Instruct-v0.2, and compare them in Examples 1, 2, and 3. As we observe, the general and detailed rationales follow our intent. Additionally, we observe that a part of the general rationale is also included in the detailed rationale. It will be important to study, which type of rationale is more helpful for the preference training.

```
<General Rationale>
The first response strictly adheres to the prompt by providing
the RDF triplets in the required format, while the second
response includes an explanation that goes beyond the required
format and may not be necessary for the task at hand.  Thus,
the first response is preferred as it focuses solely on
providing the required output in the correct format.

<Detailed Rationale>
The assistant’s response is preferred because it follows the
standard format for RDF triplets, where the subject comes
first, followed by the predicate, and then the object.  In the
given input sentence, ’AFC Ajax (amateurs)’ and ’Ajax Youth
Academy’ are the subjects, ’has ground’ and ’plays at’ are the
predicates, and ’Sportpark De Toekomst’ is the object that is
common to both subjects.  Therefore, the RDF triplets should be
structured as [subject, predicate, object], with each triplet
representing a distinct relationship.  The assistant’s response
correctly reflects this structure and semantics, making it a
preferred option.
```

Example 1: An example #1 of general and detailed rationales.

```
<General Rationale>
The preferred response strictly follows the prompt question by
calculating the total amount paid per episode for both major
and minor characters based on the given information.  It uses
the provided information to calculate the salaries of the major
and minor characters and then adds them together to find the
total amount paid per episode.  The response below includes
unnecessary details and makes the response longer and less
efficient.

<Detailed Rationale>
The assistant’s response above is clear and accurate.  They
first calculate the total amount paid to the minor characters
per episode, which is $60,000.  Then, they calculate the
total amount paid to the major characters per episode, which
is $225,000.  Finally, they add the salaries of the major
and minor characters to find the total amount James pays per
episode, which is $285,000." This explanation directly points
to the steps taken in the assistant’s response and explains
each step in detail.  It also uses clear and concise language
to make the explanation easy to understand.
```

Example 2: An example #2 of general and detailed rationales.

```
<General Rationale>
The response above is preferred because it directly answers the
question by specifically identifying the kidneys as the organs
that secrete hormones to help maintain homeostasis, while
the response below goes into excessive detail and provides
additional information that is not directly related to the
question.

<Detailed Rationale>
The assistant's response is preferred because it provides
a clear and detailed explanation of the hormones secreted
by the kidneys and their roles in maintaining homeostasis.
The response also uses the terminology from the paragraph,
such as 'erythropoietin' and 'renin,' which demonstrates a
thorough understanding of the topic.  In contrast, the less
preferred response only mentions that the kidneys secrete
hormones for homeostasis without specifying which hormones or
their functions.  Therefore, the preferred response is more
informative and accurate, making it a better answer for the
given question.
```

Example 3: An example #3 of general and detailed rationales.
winrate using HALOs repository Ethayarajh et al. (2023)

B.7 Evaluating Responses with LLM as a Judge

```
Here is a conversation between <user> and <assistant>.
<history>
<assistant> can now choose from different responses.
Response 1:  <response1>
END OF RESPONSE 1
Response 2:  <response2>
END OF RESPONSE 2
Between these two responses, which response is more helpful,
harmless, and concise?  Pick one.
```

Prompt 3: Prompt to the LLM (e.g., GPT-4o) to retrieve the preference over two responses.
After training the models with preference learning algorithms, we want to compare the performance between the models. Thus, a common practice is to generate the responses from respective models and compare the quality between them. As a human evaluator might be expensive to acquire, we usually use a capable language model to judge the responses. In our case, we leverage the GPT-4o model to judge and provide winrate scores. We provide the prompt to determine a preferred response between generations in Prompt 3, which is implemented in the HALOs code implementation (Ethayarajh et al., 2023) and shuffles responses to avoid the ordering bias.

B.8 DPO and RDPO Generated Responses

We provide some examples and compare responses generated by the models trained with DPO and RPDO losses, respectively. In Example 5, we observe the case when DPO demonstrates the verbosity and generates a lengthy output compared to the response by the RDPO model. The length bias in DPO is also demonstrated in Example 4, which generates not only a longer output than RDPO but also outputs a personal opinion on a political stance. Example 6 asks for generating a tweet for a single sentiment. While the RDPO-trained model correctly recognizes the sentiment and creates a tweet for the selected sentiment, the DPO-trained model produces tweets for all sentiments regardless of the prompt. Despite the responses generated by RDPO being shorter on average than the ones by DPO, when prompted to generate an article, the RDPO response can be as lengthy as the DPO one as shown in Example 7.

```
<Prompt>
What is your opinion on Britain exiting the EU?
What is your opinion on Britain leaving the EU and Brexit?
Would you say that these questions are the same?

<DPO Response>
These questions are related but not exactly the same.  Both
ask for an opinion on Britain's decision to leave the European
Union (EU), which is known as Brexit.  However, the first
question asks specifically about Britain exiting the EU, while
the second question expands to include the broader concept of
Brexit, which includes not only leaving the EU, but also the
negotiations and potential outcomes of that decision.

Britain exiting the EU refers to the actual act of leaving
the EU, which includes leaving the EU's political and economic
structures.  Brexit, on the other hand, encompasses the entire
process, including the negotiations, the potential outcomes,
and the impact on British politics and society.

So, if someone is asking for your opinion on Britain exiting
the EU, they are asking for your thoughts on the actual act
of leaving the EU. If they ask for your opinion on Brexit,
they might be asking for your thoughts on the entire process,
including the negotiation period and the potential outcomes of
leaving the EU.

My personal opinion is that Brexit has created uncertainty
and potential risks for both the EU and the UK. However, it
also gives the UK the opportunity to set its own trade and
regulatory policies, which could potentially lead to new
opportunities and economic growth.

<RDPO Response>
Yes, I would consider the questions to be similar, but not
exactly the same.  The first question asks for your opinion
on Britain's departure from the EU, while the second question
specifically refers to Brexit, which is the name given to the
process of Britain's departure from the EU. So, while the
second question is a subset of the first, it adds a layer of
specificity by referring to the name given to the process.
```

Example 4: A comparison of generated responses from DPO and RDPO models.

```
<Prompt>
I ' ve already decided that I want to live in Finland for a
semester .  What then ?  What if - I know , I ' m getting ahead
of my own dream - world - he and I are together for so long ...
we fall in love ?  What then ?
Answer the following question:  Where is the narrator likely to
be in a few months ?

<DPO Response>
In a few months, the narrator is likely to be living in Finland
for a semester as they have already decided on that.
As for their second question, if they fall in love with someone
during their time in Finland, the outcome could vary.  The
narrator could choose to continue their relationship despite
the distance if they are committed to making it work.  They
could also choose to pursue the relationship further if both
parties are interested.
However, it's important to note that the narrator's
hypothetical situation might not always align with reality.  If
they do fall in love and pursue a long-distance relationship,
it could require more effort, patience, and communication
to maintain the relationship.  Ultimately, it depends on the
individuals involved and their level of commitment to making it
work.
In any case, their current plan is to go to Finland for a
semester, so that is where they are likely to be in a few
months.
Regarding the world reference in the second question, it's not
clear how it relates to the overall context of the scenario.
If it refers to a fictional world, it's likely that the
question is asking about the potential outcomes of an imaginary
situation.
However, if it refers to the real world, it's worth noting that
long-distance relationships can be challenging, but they are
also possible with dedication, trust, and open communication.
If the narrator decides to pursue a relationship with someone,
they can consider using technology to stay connected, such as
video calls, messaging, and email.  They could also set aside
time for regular updates on each other's lives, and plan visits
to see each other as often as possible.
Ultimately, the key to maintaining a successful long-distance
relationship is to approach it with a positive attitude,
patience, and an understanding that challenges will arise.
By staying committed to each other and working together
to overcome any obstacles, the narrator and their partner
could potentially build a strong and lasting connection.  So,
assuming the narrator's plan to go to Finland for a semester
remains unchanged, they are likely to be in Finland in a few
months.  Whether or not they fall in love, the outcome depends
on the individuals involved and their willingness to make the
relationship work despite the challenges.

                    Response continues...

<RDPO Response>
The narrator is likely to be living in Finland for a semester.
However, if they fall in love with someone, their plans might
change depending on their feelings and circumstances.
```

Example 5: A comparison of generated responses from DPO and RDPO models.

```
<Prompt>
Sentiment possibilities.  Choices:
(A). negative;
(B). positive;.  Generate a tweet that has the following
sentiment:  (A).

<DPO Response>
(A). negative:  "I just received a faulty product from the
company, and it’s really frustrating!  #CustomerServiceFail"
(B). positive:  "Had an amazing time at the concert last night!
The energy was contagious and I’m already counting down the
days for the next one!

<RDPO Response>
Here’s a tweet that conveys a negative sentiment:
"Feeling down today.  The rain outside matches my mood.
```

Example 6: A comparison of generated responses from DPO and RDPO models.

<Prompt>
Write an article based on this summary:
Jack Tector and Nathan Smith have received their first Ireland call-up for this month's ICC Intercontinental Cup game against the Netherlands.

<DPO Response>
Title: Fresh Blood in the Irish Squad: Jack Tector and Nathan Smith Await Their Debut against the Netherlands

Ireland, a long-standing cricket force in European cricket, is set to feature two of its exciting young talents, Jack Tector and Nathan Smith, in this month's ICC Intercontinental Cup game against the Netherlands. These young guns will be making their debut on the international stage, providing them with an unforgettable experience and an opportunity to shine amongst the cricketing world's elite.

Jack Tector, a 23-year-old all-rounder, has been a consistent performer in both domestic and provincial cricket, earning him a well-deserved Ireland call-up. With his skills in both batting and bowling, Tector offers immense potential to the side and can add valuable depth to the Irish squad. His aggressive approach at the crease, coupled with his ability to generate pace and bounce from the ball, make him a formidable prospect to look out for.

Nathan Smith, 21, also makes his international debut after showcasing his prowess in various domestic competitions. He is known for his deceptive left-arm spin, which has troubled numerous batsmen in the domestic circuit.

<RDPO Response>
Ireland Calls Up Jack Tector and Nathan Smith for Upcoming ICC Intercontinental Cup Game against Netherlands

The Irish cricket team has made some exciting announcements recently, as two young and promising players, Jack Tector and Nathan Smith, have been granted their first call-ups to the national squad for the ICC Intercontinental Cup game against the Netherlands. This upcoming match is set to take place this month, and both Tector and Smith are eager to make the most of this incredible opportunity.

Jack Tector, a 21-year-old all-rounder from Wiltshire, England, has been on the radar of Irish selectors for some time now. He made his first-class debut for the Northamptonshire Steelbacks in 2019, and since then, he has been making waves in domestic cricket competitions, particularly in the 50-over format. Tector has been consistently impressive with the bat, having scored a century and several half-centuries in various domestic games. He is also a handy left-arm spinner, which adds depth to Ireland's bowling attack.

Nathan Smith, on the other hand, is a 23-year-old left-handed opening batsman from County Antrim, Northern Ireland.

Example 7: A comparison of generated responses from DPO and RDPO models.

B.9 HARDWARE AND SOFTWARE

We conduct our experiments on 3 x Nvidia A100 GPUs. Training $\leq 8B$-parameter models on preference learning with $12{,}000$ training data points requires $\sim 2-3$ GPU hours while generating responses from 512 test samples takes around 10 minutes. We publish our code implementation for generating rationales and training with rationale-enriched preference learning in the anonymous repository for the time of review: https://anonymous.4open.science/r/rationale-3973/.