HEAL: A Hypothesis-Based Preference-Aware Analysis Framework

Anonymous ACL submission

Abstract

Preference optimization methods like DPO have achieved remarkable performance in LLM alignment. However, the evaluation for these methods relies on a single response and overlooks other potential outputs, which could also be generated in real-world applications within this hypothetical space. To address this issue, this paper presents a **H**ypothesis-based Pr**E**ference-aware **A**na**L**ysis Framework (HEAL), a novel evaluation paradigm that formulates preference alignment as a re-ranking process within hypothesis spaces. The framework incorporates two complementary metrics: ranking accuracy for evaluating ordinal consistency and preference strength correlation for assessing continuous alignment. To facilitate this framework, we develop UniHypoBench, a unified hypothesis benchmark constructed from diverse instruction-response pairs. Through extensive experiments based on HEAL, with a particular focus on the intrinsic mechanisms of preference learning, we demonstrate that current preference learning methods can effectively capture preferences provided by proxy models while simultaneously suppressing negative samples. These findings contribute to preference learning research through two significant avenues. Theoretically, we introduce hypothesis space analysis as an innovative paradigm for understanding preference alignment. Practically, HEAL offers researchers robust diagnostic tools for refining preference optimization methods, while our empirical results identify promising directions for developing more advanced alignment algorithms capable of comprehensive preference capture.

1 Introduction

Direct preference optimization (DPO) has emerged as the predominant method for aligning large language models (LLMs) with human preferences (Rafailov et al., 2024). Recent research on DPO has also explored various variants, including SimPO (Meng et al., 2024), ORPO (Hong et al., 2024), and KTO (Ethayarajh et al., 2024). To evaluate the effectiveness of these preference alignment methods, researchers typically rely on downstream benchmarks such as AlpacaEval (Dubois et al., 2024) and MT-Bench (Bai et al., 2024). In the evaluation process using these benchmarks, we typically follow a standard procedure: given a prompt, we first generate a response from the aligned model using a temperature-based sampling method (Gu et al., 2024). Next, we employ a proxy model (such as GPT-4) to compare this response with the reference response, evaluating whether the model is effectively aligned.

However, this procedure faces a fundamental limitation because this *sampling-based* evaluation approach only examines single responses sampled from target LLMs. But in practice, the LLMs are commonly expected to generate a wide spectrum of diverse responses, which cannot be sufficiently covered by several sampled responses. This misalignment between evaluation and real-world LLM development prohibits researchers and developers from assessing the LLMs' performance accurately. Furthermore, this paradigm also neglects the relative comparison of responses, which is fundamentally modeled in preference learning scenarios.

To address these limitations in the evaluation, we propose HEAL (**H**ypothesis-based pr**E**ference-aware **A**na**L**ysis), a novel framework that evaluates LLMs through the lens of ranked hypothesis spaces. Inspired by the ranking-based evaluation approaches such as RewardBench (Lambert et al., 2024), HEAL conceptualized preference alignment as a dynamic re-ranking process within the hypothesis space, enabling comprehensive assessment through two complementary quantitative methods: (1) The first metric is ranking accuracy, which is measured via Kendall's Tau between the policy model's rankings and proxy preference model rankings (used for training data annotation). This

metric evaluates ordinal consistency in preference learning and directly assesses whether the model preserves the relative ordering of hypotheses as intended by the preference signals. (2) The second one is preference strength correlation. We use Pearson correlation between generation likelihoods and gold-standard preference scores as the metric. This metric captures continuous alignment precision. Unlike binary or ordinal measures, it quantifies the model's sensitivity to fine-grained preference distinctions, ensuring quantified relationships in preference strength are maintained across hypotheses.

We evaluate mainstream preference learning methods using HEAL to address three key research questions. (RQ1): Do these methods effectively capture preference information? While ranking accuracy confirms that LLMs acquire preferences through optimization, results reveal incomplete learning. To elaborate, current methods struggle to fully absorb all preference signals. (RQ2): Can LLMs discern proxy model-specific preferences? Experiments demonstrate that LLMs successfully learn distinct preference patterns from different proxy models, showcasing HEAL's sensitivity to subtle inter-model variations. (RQ3): How do learned preferences vary across methods? All tested methods achieve strong in-distribution alignment with proxy models, but out-of-distribution performance degrades significantly, except for SimPO, which exhibits notable generalization. These results validate the partial efficacy of preference alignment while underscoring critical limitations, particularly in robustness and completeness of learned preferences.

Our main contributions are:

- To the best of our knowledge, we are the first to present a systematic study assessing how effectively LLMs capture proxy model preferences through a hypothesis lens.
- We construct HEAL, a hypothesis-based preference-aware analysis framework that quantifies the preference modeling analysis into two metrics: ranking accuracy and preference strength correlation. Furthermore, we construct a **Uni**fied **Hypo**thesis **Bench**mark (UniHypoBench) to support the evaluation pipeline of HEAL.
- Comprehensive experiments demonstrating HEAL's effectiveness while revealing key limitations in current preference learning methods, particularly regarding robustness and preference completeness.

2 Preliminary

2.1 Sequence Likelihood

In the literature of LLMs, a model parameterized with θ is essentially a generative large language model, which is applied to generate a response sequence y when prompted with input x. The response y is typically generated by sampling the next tokens auto-regressively from a probabilistic distribution. At each time step in this procedure, the model selects the next token randomly to form a new input. Under this approach, the likelihood of generating a specific sentence can be obtained by computing the conditional probability, which can be written as:

$$\pi_\theta(y|x) = \prod_{n=0}^{|y|} P_\theta(y_n|y_{<n}, x) \quad (1)$$

where the term $P_\theta(y_n|y_{<n}, x)$ represents the probability of the $n-\text{th}$ token of response y. The sequence likelihood reflects how an LLM tends to generate a specific response, and this likelihood also serves as a core component of other metrics, such as perplexity (PPL).

2.2 Human Preference Alignment

In the realm of LLMs, the training process generally encompasses three key stages: pre-training, supervised fine-tuning, and human preference alignment (Ouyang et al., 2022). Recent advancements have demonstrated that alignment can be effectively achieved through two branches of RLHF: reward-based methods and reward-free methods. Throughout the training process, LLMs inherently learn human preferences either through reward scores provided by a reward model or by utilizing ranked pairs of responses. However, the reward-based methods often require extensive reward modeling and face challenges in scalability and generalization (Gao et al., 2022). Consequently, recent work has shifted towards reward-free methods, which directly optimize preferences without explicit reward signals. This shift highlights the growing importance of reward-free approaches in addressing the limitations of traditional reward-based methods, offering a more scalable and efficient path for aligning LLMs with human preferences.

Reward Modeling. To effectively capture human preferences, a widely adopted approach involves training a reward model using human preference datasets. In the context of RLHF, a reward model is generally formulated as a function $r_\phi(x, y)$, where ϕ represents model parameters, x denotes the instruction, and y corresponds to the response. To develop such a reward model, a foundation LLM is optimized by minimizing the Bradley-Terry loss (Stiennon et al., 2020), as follows:

$$\mathcal{L}_{\text{reward}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_\text{p}} \log(\sigma(r_\phi(x, y_w) - r_\phi(x, y_l))) \quad (2)$$

Here, D_p represents the human preference dataset, which comprises input tuples containing an instruction x and a pair of responses (y_w, y_l) with preference $y_w \succ y_l$, where $y_w \succ y_l$ indicates that y_w is preferred over y_l according to human or model-based annotations. This dataset serves as the foundation for training the reward model, enabling it to capture human preferences effectively.

Although it has been discussed that recent work has increasingly focused on reward-free methods, reward models continue to play a significant role in alignment. For instance, a robust reward model can act as a reliable human proxy, which is capable of constructing high-quality preference data for reward-free methods such as DPO. The training of reward models lays the groundwork for understanding and optimizing human preferences, which will be further explored in the context of preference optimization in subsequent sections.

Preference Optimization. Building on the discussion of reward models and their role in alignment, Rafailov et al. (Rafailov et al., 2024) introduced DPO, a novel approach that inherently integrates the reward model within the policy model itself. In DPO, the policy model is directly optimized using a preference dataset, which can also serve as the basis for training a reward model. This dual-purpose utilization of the dataset highlights the flexibility and efficiency of the DPO. The DPO loss function can be given by:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_\text{p}} \left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \quad (3)$$

where β denotes a parameter that controls the strength of constraints, ensuring the optimized policy $\pi_\theta(y|x)$ does not deviate excessively from the reference policy $\pi_{ref}(y|x)$. Notably, there is a significant conceptual similarity between the loss functions of DPO and reward modeling loss as introduced in Eq. 2. Although the specific objectives of Eq. 2 and Eq. 3 differ in formulation, their underlying goals are fundamentally aligned: *Maximize the likelihood of generating responses preferred by humans while minimizing the probability of producing dispreferred ones*. In conclusion, understanding the shared principles of preference modeling between reward-based and reward-free methods is crucial for uncovering the fundamental mechanisms of alignment in LLMs.

3 HEAL: A Hypothesis-based Analysis Framework

We propose HEAL (hypothesis-based analysis), illustrated in Figure 1. The framework models preference patterns as ranked hypothesis spaces and evaluates them through two complementary metrics: (1) ranking accuracy for ordinal consistency and (2) preference strength correlation for continuous alignment, which we detail in this section respectively.

3.1 Ranking in Hypothesis Space

In the current evaluation procedure, recent research studies the behavior of LLMs directly from their generated content. However, as mentioned in Sec. 2.1, generation is naturally a random process that hardly produces stable outputs.

Definition 1 (*Hypothesis Space*)**.** In the generation phase, the responses possibly differ within a constrained set Y_x due to variations in hyperparameter configurations. To study these responses, we extend the term hypothesis to LLMs by analogy to natural language understanding (NLU), where this terminology denotes a candidate sentence sampled from the output space (Proebsting and Poliak, 2024). Here, we consider the constrained set Y_x as the hypothesis space for the given x, which contains all possible responses (*i.e.* hypotheses). The hypothesis space is formulated as follows:

$$Y_x = \{y_i \in \bigcup_{n=0}^{\infty} V^n \,|\, \mathbf{I}(x, y_1) \geq \cdots \geq \mathbf{I}(x, y_i) \geq \cdots, i \in N_+\} \quad (4)$$

where V denotes the vocabulary set defining the hypothesis space. Formally, Y_x constitutes a set comprising (potentially infinite) textual hypotheses

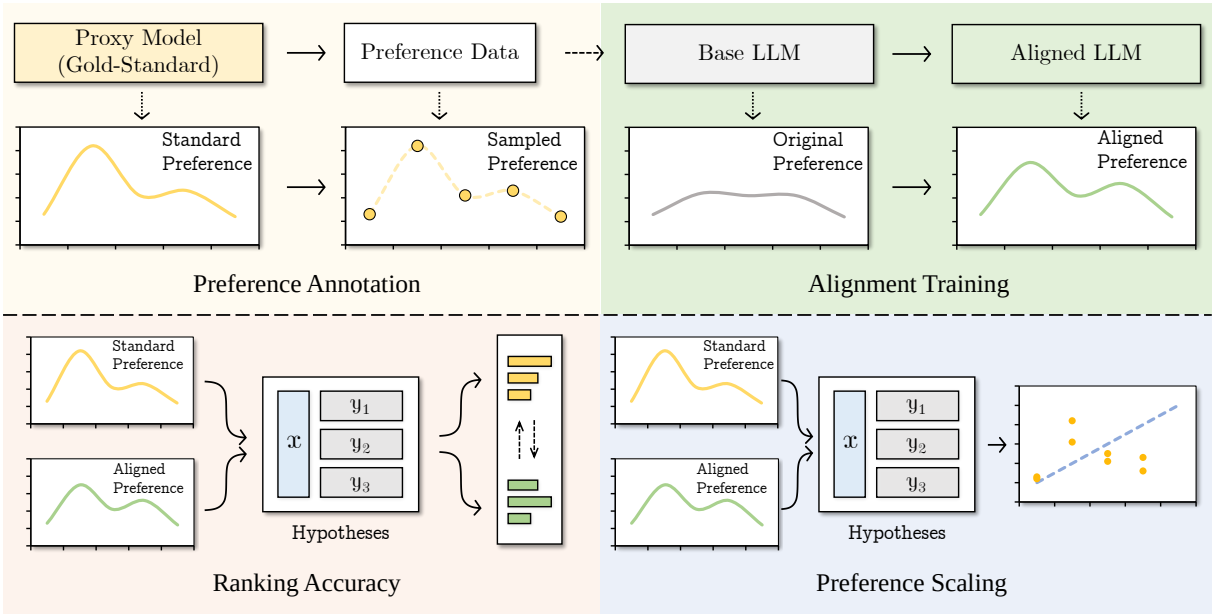


Figure 1: The overview of HEAL. We conceptualize preference learning as an alignment process between the original preference pattern and the standard preference sampled from the proxy model. The framework employs two evaluation metrics - ranking accuracy and preference strength correlation.

generated from V. These hypotheses are ordered by an indicator function $\mathbf{I}(x, y)$, which assigns a comparable scalar value to each hypothesis y. Typical instantiations of $\mathbf{I}(x, y)$ include the generation probability $\pi_\theta(y|x)$ under the model parameter θ, or a preference score given by a human annotator (or a strong LLM). With these instantiations, the hypothesis space Y_x is structured as an ordered set, where hypotheses are ranked in descending order. Consequently, hypotheses positioned earlier in Y_x exhibit a higher likelihood of being selected during generation.

Definition 2 (*Gold-Standard Hypothesis Space*)**.** From the perspective of generation probabilities, the alignment algorithms optimize the generation probabilities to favor preferred responses while suppressing the dispreferred ones, thus we can conclude alignment as reordering the hypothesis space Y_x to better match a gold-standard hypothesis space $Y_{\text{x;gold}}$. Formally, the gold-standard hypothesis space $Y_{\text{x;gold}}$ is defined as:

$$Y_{\text{x;gold}} = \{y_i \in \bigcup_{n=0}^{\infty} V^n \,|\, \text{GS}(x, y_1) \geq \cdots \geq \text{GS}(x, y_i) \geq \cdots, i \in N_+\} \quad (5)$$

where the gold scoring function $\text{GS}(x, y)$ quantifies the alignment quality of response y to instruction x, as evaluated by either reward models or human annotators (Lambert et al., 2024; Zhou et al., 2023). From a formal perspective, $\text{GS}(x, y)$ represents a specialized instantiation of the indicator function $\mathbf{I}(x, y)$, optimized to reflect ideal human preferences. In conclusion, this space serves as the theoretical optimum for alignment objectives.

3.2 Quantitative Analysis Method

Ranking Accuracy. In Section 3.1, we formally defined both the hypothesis space Y_x and the gold-standard hypothesis space $Y_{\text{x;gold}}$. These spaces differ only in their internal ranking criteria, while their elements remain identical (Chen et al., 2024).

To quantify the alignment quality between an LLM's outputs and gold-standard preferences, we propose measuring the ordinal discrepancy between Y_x and $Y_{\text{x;gold}}$. Specifically, we adapt Kendall's Tau-b correlation coefficient as our metric for comparing their partial orders. The Kendall's Tau-b statistic is formally expressed as:

$$\tau_b(Y_x^{(1)}, Y_x^{(2)}) = \frac{\text{C}(Y_x^{(1)}, Y_x^{(2)}) - \text{D}(Y_x^{(1)}, Y_x^{(2)})}{\sqrt{(\text{T}_0 - \text{T}_1(Y_x^{(1)}))(\text{T}_0 - \text{T}_2(Y_x^{(2)}))}} \quad (6)$$

where:

- $Y_x^{(1)}$ and $Y_x^{(2)}$ denote two hypothesis spaces sharing identical response elements but potentially differently ordered by their respective indicator functions $\mathbf{I}_1(x, y)$ and $\mathbf{I}_2(x, y)$.

- $\mathrm{C}(\cdot,\cdot)$ counts concordant pairs - cases where the relative ordering of (y_i, y_j) is consistent between both spaces, while $\mathrm{D}(\cdot,\cdot)$ counts discordant pairs with contradictory orderings.
- $\mathrm{T}_0 = \binom{n}{2}$ represents the total possible pairs.
- $\mathrm{T}_1(\cdot)$ and $\mathrm{T}_2(\cdot)$ are tie correction terms for each hypothesis space.

The ranking of responses y_i, y_j in each space is determined by comparing their indicator values $\mathbf{I}(x, y_i)$ and $\mathbf{I}(x, y_j)$. The denominator's adjustment for ties ensures robustness when the indicator function produces discrete scores. This metric provides a unified comparison capability, applicable to any system generating comparable $\mathbf{I}(x, y_i)$ values.

Definition 3 (*Ranking Accuracy*)**.** Furthermore, we notice that Kendall's Tau differs from the ranking accuracy in range. Therefore, we map the original metric to obtain an accuracy ratio, as follows:

$$\mathrm{RA}(\mathcal{D}) = \mathbb{E}_{(x,Y_x^{(1)},Y_x^{(2)})\sim\mathcal{D}} \frac{\tau_b(Y_x^{(1)}, Y_x^{(2)}) + 1}{2} \quad (7)$$

where the term $\tau_b(\cdot,\cdot)$ is computed based on Eq. (6). Here, $\mathcal{D}$ denotes the input dataset consisting of tuples $(x, Y_x^{(1)}, Y_x^{(2)})$. The mapping operation is equal to assigning a zero-valued weight to the discordant pairs since they do not contribute to the accuracy computation.

Preference Strength Correlation. In human preference modeling, beyond relative ranking, preference strength correlation plays a critical role by quantifying the strength of preferences through continuous numerical values. However, the current reward-free alignment paradigm often disregards this scalar information, focusing solely on ordinal comparisons. This omission leads to a loss of preference modeling, which may result in LLMs that fail to accurately capture the subtle variance in human preferences. Consequently, such models can exhibit suboptimal calibration in generation likelihoods or reward predictions (Zhou et al., 2024).

Definition 4 (*Preference Strength Correlation*)**.** We consider that if an LLM is perfectly aligned with a gold-standard hypothesis space, its indicator function values should exhibit a strong linear correlation with those derived from the gold standard. To quantify the correlation at the dataset level, we propose an expectation-based Pearson correlation metric, as follows:

$$\mathrm{PSC}(\mathcal{D}) = \mathbb{E}_{(x,Y_x^{(1)},Y_x^{(2)})\sim\mathcal{D}} \left[\frac{\mathbb{E}[I_1 I_2] - \mathbb{E}[I_1]\mathbb{E}[I_2]}{\sigma_{I_1}\sigma_{I_2}} \right] \quad (8)$$

where the function $\mathbf{I}_1(\cdot,\cdot)$ and $\mathbf{I}_2(\cdot,\cdot)$ represent the indicator functions (generation likelihoods or reward scores) of hypothesis spaces $Y_x^{(1)}$ and $Y_x^{(2)}$ respectively. Consequently, a well-aligned model would yield a Pearson correlation coefficient approaching 1, reflecting high agreement in preference strength correlation.

4 Experiments

4.1 Setups

We evaluated three widely adopted preference optimization algorithms, including DPO (Rafailov et al., 2024), SimPO (Meng et al., 2024), and ORPO (Hong et al., 2024), using our proposed framework. For preference annotation and evaluation, we employed ArmoRM-LLaMA-3-8B-v0.1 (Wang et al., 2024a) as our primary gold-standard proxy model, ensuring consistency between training and evaluation preference distributions. To investigate the influence of optimization methods across different preference distributions, we additionally utilized GRM-LLaMA3-8B-rewardmodel-ft (Yang et al., 2024) as a comparative proxy model with a distinct preference distribution.

Datasets. We employed the following datasets for training and evaluation:

(1) **UltraFeedback** (Cui et al., 2023): A large-scale preference dataset comprising 64k prompts and 256k responses. We performed preference optimization on the training split and utilized its validation set for in-distribution evaluation.

(2) **HelpSteer2-Preference** (Wang et al., 2024b): A high-quality dataset annotated with preference directions, strength scores, and textual justifications. Similarly, we conducted the evaluation on its validation split.

(3) **UniHypoBench**: To address the limitation of the existing evaluation sets (which typically provide less than 4 responses per prompt), we constructed the Unified Hypothesis Benchmark (UniHypoBench), as detailed in Appendix A.1. Curated from RewardBench (Lambert et al., 2024), it extends the evaluation scope with 2,985 prompts, each containing more than 8 responses sampled

Model/Method	w/o Length Normalization						w/ Length Normalization					
	UniHypo		HelpSteer2		UltraFeedback		UniHypo		HelpSteer2		UltraFeedback	
	RA	PSC	RA	PSC	RA	PSC	RA	PSC	RA	PSC	RA	PSC
Alignment with ArmoRM-Llama3-8B-v0.1 (Same Preference Distribution)												
LLaMA-3.2-3B-Instruct	54.64	0.152	**46.79**	**-0.063**	53.09	0.079	48.22	-0.048	50.69	0.013	**49.44**	-0.017
+DPO	**54.72**	**0.154**	**46.79**	**-0.063**	53.17	**0.081**	**48.29**	**-0.046**	**51.38**	**0.027**	49.38	**-0.016**
+ORPO	54.55	0.151	46.68	-0.065	53.12	0.080	48.19	-0.049	50.69	0.013	49.28	-0.018
+SimPO	54.61	0.152	46.68	-0.066	**53.22**	0.080	48.21	-0.048	51.16	0.023	49.31	-0.017
LLaMA-3-8B-Instruct	54.15	0.124	47.36	-0.051	53.13	0.079	49.81	0.031	50.81	0.016	49.86	-0.011
+DPO	54.16	0.138	49.31	-0.013	64.62	0.368	47.66	-0.033	55.76	0.112	59.87	0.251
+ORPO	52.67	0.084	48.39	-0.031	63.28	0.341	48.37	-0.029	53.33	0.065	64.45	0.065
+SimPO	**63.59**	**0.502**	**53.32**	**0.065**	**66.70**	**0.419**	**73.30**	**0.598**	**66.51**	**0.319**	**71.51**	**0.545**
Alignment with GRM-Llama3-8B-rewardmodel-ft (Different Preference Distribution)												
LLaMA-3.2-3B-Instruct	**51.65**	**0.059**	**53.14**	**0.063**	50.78	0.021	**52.08**	**0.066**	48.42	-0.031	51.16	0.030
+DPO	51.64	**0.059**	52.91	0.058	50.77	0.021	**52.08**	**0.066**	**48.65**	**-0.027**	**51.29**	0.031
+ORPO	**51.65**	**0.059**	**53.14**	**0.063**	50.74	0.021	**52.08**	**0.066**	**48.65**	**-0.027**	51.21	0.032
+SimPO	**51.65**	**0.059**	**53.14**	**0.063**	**50.84**	**0.022**	**52.08**	**0.066**	48.19	-0.036	**51.29**	**0.033**
LLaMA-3-8B-Instruct	**51.68**	**0.056**	52.58	0.051	49.92	0.008	**52.67**	**0.076**	48.53	-0.029	50.13	0.013
+DPO	51.29	0.049	54.48	0.089	**50.54**	**0.014**	51.78	0.054	47.52	-0.049	50.31	0.007
+ORPO	51.56	0.053	52.47	0.049	50.10	0.009	52.42	0.072	**49.66**	**-0.007**	50.01	0.007
+SimPO	49.50	-0.031	**55.03**	**0.100**	50.46	0.013	50.73	-0.008	48.98	-0.020	**50.99**	**0.023**

Table 1: Experimental results on different preference optimization methods. RA and PSC denote ranking accuracy and preference strength correlation, respectively. The best results for each group are in **bold**. The second-best results for each group are with underline.

from diverse commercial and open-source LLMs, enabling more comprehensive analysis.

Models. We evaluated our approach using three models, including LLaMA-3.2-3B-Instruct and LLaMA-3-8B-Instruct. For the LLaMA-3.2-3B-Instruct model base, we conducted preference optimizations. Concurrently, the LLaMA-3-8B-Instruct models were evaluated using the pre-optimized weights released by Meng et al. (2024).

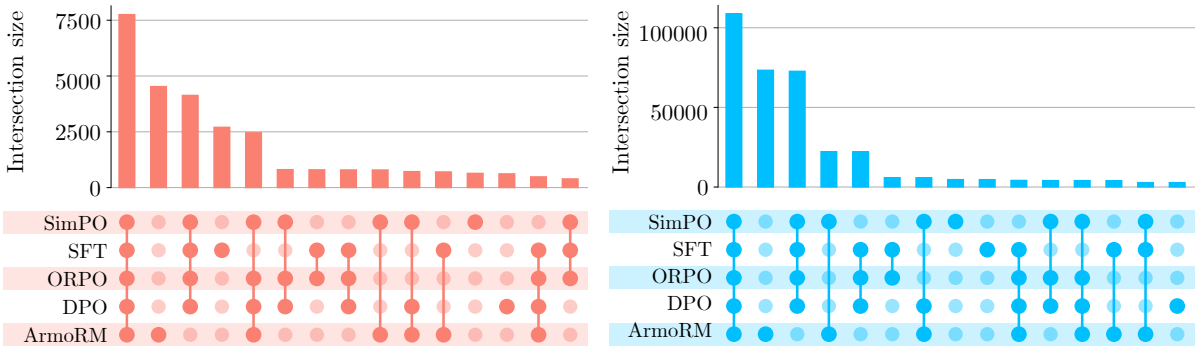
Training Settings. We conducted preference optimization using an effective batch size of 128 and a maximum sequence length of 1024. The learning rate follows a cosine decay schedule with 10% warmup steps over one training epoch. For method-specific hyperparameters, we performed a grid search to determine the optimal configuration.

4.2 Main Results

We conduct an evaluation of diverse preference learning methods using our proposed framework. To ensure the preference consistency, we maintain identical rating proxy models between training and evaluation phases, thereby guaranteeing that all compared methods learn from and are assessed against the same preference distribution. Additionally, we establish the original instruction-tuned models as baseline comparisons, which enables a quantitative assessment of the performance gains achieved through explicit preference learning. We present the main results in Table 1. The results demonstrate:

Preference Optimization Effectively Captures the Preference Information. Our results show that preference optimization methods generally outperform baselines in both ranking accuracy and preference strength correlation, confirming their effectiveness in capturing preference distributions. The LLaMA-3-8B-Instruct model benefits most significantly, with SimPO achieving over 10% improvement across all datasets. However, even SimPO's best performance (66.70% on UltraFeedback) remains suboptimal, aligning with Chen et al.'s observation that current methods still have substantial room for improvement.

Preference Optimization Learns Model-Specific Preference Patterns. Our evaluation reveals that while preference optimization improves alignment with the training proxy model, these gains often fail to generalize to other proxy models with different distributions. In some cases, we even observe performance degradation when evaluating against alternative proxies. These findings demonstrate that current methods primarily learn model-specific judging patterns rather than general preferences.



(a) ID Preference Overlaps (b) OOD Preference Overlaps

Figure 2: Upset plots of preference intersections on the OOD test set (UniHypo). The upper bar chart displays the amount of preference overlaps between different methods, while the lower connection matrix identifies the constituent subsets of each intersection. Full results can be found in Figure 5.

This specificity poses a fundamental challenge for the LLM-based evaluation, as different evaluators may employ conflicting preference criteria, complicating the assessment of alignment quality.

Length Normalization is Potential in Preference Modeling. Our analysis indicates that length normalization generally impairs ranking accuracy. However, the normalized SimPO version of LLaMA-3-8B-Instruct achieves a 73.3% accuracy, surpassing the performance of unnormalized models. This demonstrates that length-aware objectives can learn better preference representations, suggesting their value for future methods.

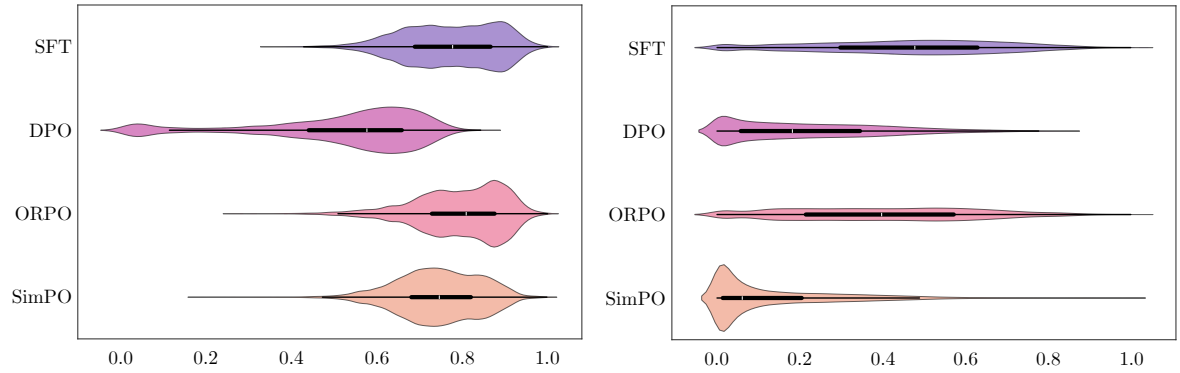
Capturing Subtle Preference Correlation is Challenging. Current alignment methods exhibit strong ranking accuracy but exhibit a weak correlation with preference strength, typically below 0.3. This result highlights the challenges in quantifying preference strength. However, SimPO stands out as an exception, achieving a preference strength correlation of 0.419 on UltraFeedback (up from 0.079), demonstrating that improved strength modeling is achievable.

4.3 Analysis

Visualization of Preference Intersections. We employ upset plots (Lex et al., 2014) to analyze preference intersections across hypothesis spaces (Figure 2), presenting both in-distribution (a) and out-of-distribution (b) results. For clarity, we focus on the plot's forepart, where solid-connected points mark shared preference tendencies across response pairs. Our key observations of the in-distribution test set include: (1) The largest intersection demonstrates fundamental preference knowledge shared by all optimization methods and the proxy model. (2) The substantial second intersection indicates significant unlearned preferences. (3) The fourth intersection shows that methods successfully capture novel preferences from the proxy model with notable behavioral deviation.

In parallel, we conduct the identical visualization on the out-of-distribution test set. Apart from the observations in Figure 2 (a), we find that: (1) Unlearned preferences increase proportionally, revealing domain-shift effects. (2) While most methods degrade, SimPO maintains the largest intersection, demonstrating superior generalization. (3) The overall performance decline underscores the need for more robust preference learning paradigms. These visual analyses provide intuitive mechanistic insights that corroborate our quantitative findings in Section 4.2.

Alignment in LLM's Internal Preference Distribution. To gain deeper insights into the alignment effects, we analyze the internal preference distribution using UniHypoBench. We sample responses from aligned LLaMA-3-8B-Instruct-based models (with the SFT base model as baseline) at a temperature of 0.75 to ensure sufficient diversity. The generation likelihoods are then extracted to compute both ranking accuracy and preference strength correlation, as listed in Table 2. Surprisingly, the results reveal that performance shows no significant improvement even when evaluated on the model's own preference distribution. More notably, we observe performance degradation in some cases, particularly for the SimPO-aligned model. We assume that this phenomenon probably stems from the scarcity of the diversity of these



(a) Internal Preference Distribution (b) External Preference Distribution

Figure 3: Violin plots comparing generation likelihood distributions across different alignment methods (including the SFT base model as baseline). The plot width represents probability density, with the central white line marking the median value for each distribution. To enhance the readability of the chart, we employ length normalization.

Metric/Method	SFT	DPO	ORPO	SimPO
RA w/o LN	55.93	55.99	55.01	**61.50**
PSC w/o LN	0.164	0.153	0.123	**0.301**
RA w/ LN	54.12	51.49	54.74	**59.70**
PSC w/ LN	0.100	0.044	0.108	**0.252**

Table 2: Experimental results of LLaMA-3-8B-Instruct-based models' internal preference distribution. The best results for each group are in **bold**. LN denotes length normalization.

sampled hypotheses. This finding also notes that distinguishing the subtle difference between similar hypotheses is a challenge for further development of the preference learning method.

To gain deeper insight into the mechanisms of preference learning, we conduct a density-based analysis of generation likelihood distributions. Figure 3 presents violin plots of these distributions across different alignment approaches, with the density curves estimated using kernel density methods. As shown in Figure 3 (a), which visualizes the internal preference distributions, we observe remarkably consistent patterns across all optimization methods. The finding indicates that these optimization methods probably do not vary much in their internal preference distributions.

Our analysis of external preference distributions in Figure 3 (b) reveals distinct patterns that contrast with the internal consistency observed previously. The distributions exhibit a clear dichotomy: while SFT and ORPO maintain near-uniform distributions, DPO and SimPO obtain spindle-shaped distributions, reflecting their enhanced capability to suppress negative samples through preference optimization. This successful suppression of undesirable outputs represents a significant advancement in alignment techniques.

However, closer examination reveals a critical limitation - none of the methods achieve the theoretically optimal bimodal distribution that would fully separate preferred and rejected responses. This persistent unimodality suggests that while current approaches can effectively downweight negative samples, they struggle to develop truly discriminative representations that clearly partition the hypothesis space. The gap between empirical results and theoretical expectations points to fundamental constraints in existing optimization frameworks, which appear to learn primarily through global likelihood adjustment rather than developing more sophisticated, robust representations of preference structure.

5 Conclusion

In this paper, we have explored evaluation and analysis methods for preference learning via preference-aware evaluation. Specifically, we first developed a hypothesis-based analysis framework containing two complementary metrics, HEAL. Based on HEAL, we then evaluate how effectively the LLMs capture preferences through preference learning. Furthermore, we construct UniHypoBench to support our evaluation pipeline. Extensive experiments demonstrate the effectiveness of our evaluation and analysis methods.

Limitations

While our proposed HEAL framework provides a novel hypothesis-based approach for preference-aware analysis, several limitations require discussion. First, our experimental validation, though demonstrating practical utility for resource-constrained scenarios, was conducted on a limited set of models, with LLaMA-3-8B-Instruct serving as the primary exemplar due to its consistently strong performance. Second, while ranking accuracy and preference strength correlation prove effective as evaluation metrics, future work may identify more sophisticated measures that better capture the nuances of preference learning. Finally, our current analysis does not examine the training dynamics of these metrics during optimization, leaving open questions about their evolution and relationship to model convergence. These limitations point to valuable directions for future research, particularly in developing more comprehensive analysis approaches and investigating the inherent mechanism of preference alignment.

Ethics Statement

This work does not need ethical considerations. Although in this work we construct data as described in Appendix A.1, this input is all from open-source data, and the output is also obtained based on open-source or commercial models.

References

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *ArXiv preprint*, abs/2402.14762.

Angelica Chen, Sadhika Malladi, Lily H. Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. 2024. Preference learning algorithms do not learn preference rankings. *ArXiv*, abs/2405.19534.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *ArXiv preprint*, abs/2310.01377.

Yann Dubois, Bal'azs Galambosi, Percy Liang, and Tatsunori Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *ArXiv preprint*, abs/2404.04475.

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

Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *ArXiv*, abs/2411.15594.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Haotong Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Proceedings of the 29th Symposium on Operating Systems Principles*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *ArXiv preprint*, abs/2403.13787.

Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis)*, 20(12):1983–1992.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155.

Grace Proebsting and Adam Poliak. 2024. Hypothesis-only biases in large language model-elicited natural language inference. *ArXiv*, abs/2410.08996.

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

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. ArXiv.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan J. Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *ArXiv preprint*, abs/2009.01325.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Conference on Empirical Methods in Natural Language Processing*.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024b. Helpsteer2-preference: Complementing ratings with preferences. *ArXiv*, abs/2410.01257.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024. Regularizing hidden states enables learning generalizable reward model for llms. *ArXiv*, abs/2406.10216.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *ArXiv*, abs/2403.13372.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, L. Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *ArXiv*, abs/2305.11206.

Hang Zhou, Chenglong Wang, Yimin Hu, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2024. Prior constraints-based reward model training for aligning large language models. In *China National Conference on Chinese Computational Linguistics*, pages 555–570. Springer.

Name	Commercial	Open-source
Claude-3-Haiku	✓	✗
GPT-4o	✓	✗
DeepSeek-V2-Lite	✗	✓
Qwen2.5-32B	✗	✓
Qwen-14B	✗	✓
Mixtral-8x7B-Instruct-v0.1	✗	✓
LLaMA-3-8B-Instruct	✗	✓
ChatGLM3-6B	✗	✓

Table 3: Models selected for UniHypoBench construction.

A Implementation Details

A.1 Construction of UniHypoBench

We construct UniHypoBench based on the RewardBench instruction set, leveraging its comprehensive coverage of diverse task types. Our benchmark construction process begins by collecting hypothesis samples from multiple powerful commercial and open-source LLMs, as specified in Table 3.

To enhance the response diversity while maintaining quality, we configured the sampling parameters with a temperature setting of 0.75 and top-p value of 0.95, with all responses truncated at 768 tokens. Following generation, we implemented a filtering process to remove low-quality and empty responses, thereby ensuring the benchmark's reliability and consistency.

A.2 Experimental Setups

Our implementation leverages `LLaMA-Factory` (Zheng et al., 2024) for model training and `vLLM` (Kwon et al., 2023) for efficient inference. All experiments were conducted on 2×NVIDIA 3090 GPUs, with additional optimization through `DeepSpeed` (Rajbhandari et al., 2020) ZeRO-2 to minimize computational overhead and accelerate training. Following established practices in preference optimization (Meng et al., 2024), we maintain an effective batch size of 128 and employed a cosine learning rate schedule with 10% warmup steps. To balance computational efficiency with model performance, we set the training sequence length to 1024 tokens.

Before final model training, we performed extensive hyperparameter tuning to identify optimal configurations for each method. We first search the learning rates invidually in the range of [3e-7, 7e-7, 1e-6]. Then we search method-specific parameters whose search ranges are detailed in Table 4.

B More Analysis

Preference Learning Achieves Limited Improvements with Confident LLMs. Building upon the main results presented in Table 1, we observe that preference optimization yields limited improvement for LLaMA-3.2-3B-Instruct, with both ranking accuracy and preference strength showing marginal gains or even performance degradation. This unexpected outcome suggests potential overfitting to the original training corpus during earlier optimization stages. More fundamentally, these findings reveal an important relationship between a base model's core capabilities and its capacity for effective preference learning - implying that successful alignment may be constrained by the underlying model's basic capabilities before fine-tuning.

More Results of the Main Experiment To further validate our findings, we extend the evaluation to Mistral-7B-Instruct, a widely adopted foundation model in contemporary LLM research, utilizing the optimized weights provided by Meng et al. (2024). As evidenced in Table 5, the experimental outcomes exhibit some divergence from our primary results. We hypothesize that these discrepancies stem from fundamental differences in both the base model architecture and the composition of the training corpus, highlighting the model-dependent nature of preference optimization efficacy.

C More Upset Plots

This section presents the complete upset plot visualizations in Figure 4, along with their length-normalized counterparts in Figure 5. The observed patterns remain consistent with our preliminary analysis in Figure 2, further validating our earlier conclusions regarding preference alignment behaviors. Notably, we find that the length normalization has smoothed the distribution of the intersections, which could be valuable for further study.

D Joint Distribution of Reward Scores and Generation Likelihoods

Figure 6 presents the joint distribution of reward scores and generation likelihoods, revealing several key insights about preference learning dynamics. Consistent with our previous observations, all examined methods demonstrate the capability to effectively suppress likelihoods for undesired responses, confirming this as a fundamental mecha-

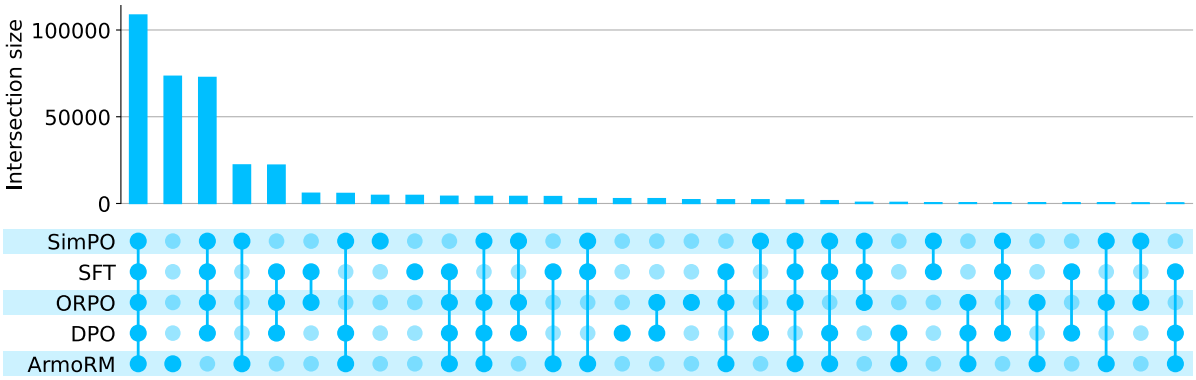
Method	Objective	Hyperparameter
DPO	$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$	$\beta\in[0.01,0.05,0.1]$
ORPO	$-\log p_\theta(y_w\|x)-\lambda\log\sigma\left(\log\frac{p_\theta(y_w\|x)}{1-p_\theta(y_w\|x)}-\log\frac{p_\theta(y_l\|x)}{1-p_\theta(y_l\|x)}\right)$, where $p_\theta(y\|x)=\exp\left(\frac{1}{\|y\|}\log\pi_\theta(y\|x)\right)$	$\lambda\in[0.1,0.5,1.0]$
SimPO	$-\log\sigma\left(\frac{\beta}{\|y_w\|}\log\pi_\theta(y_w\|x)-\frac{\beta}{\|y_l\|}\log\pi_\theta(y_l\|x)-\gamma\right)$	$\beta\in[2.0,2.5,3.0,5.0,10.0]$, $\gamma\in[0.3,0.5,1.0]$

Table 4: Optimization objectives and hyperparameter search ranges of applied preference learning methods

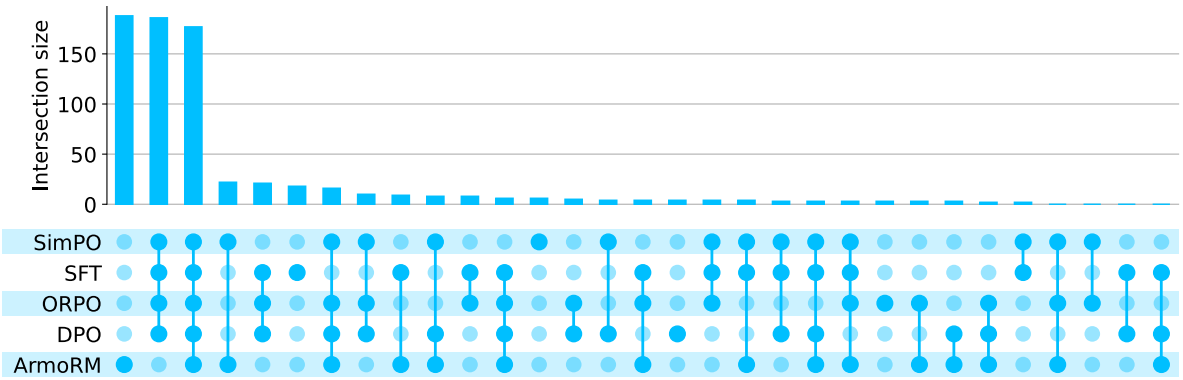
Model/Method	w/o Length Normalization						w/ Length Normalization					
	UniHypo		HelpSteer2		UltraFeedback		UniHypo		HelpSteer2		UltraFeedback	
	RA	PSC	RA	PSC	RA	PSC	RA	PSC	RA	PSC	RA	PSC
Alignment with ArmoRM-Llama3-8B-v0.1 (Different Preference Distribution)												
Mistral-7B-Instruct	62.30	0.356	52.52	0.049	54.36	0.117	54.40	0.135	55.17	0.100	50.17	0.009
+DPO	**65.99**	**0.489**	**53.79**	**0.074**	**56.27**	**0.165**	**58.45**	**0.202**	57.01	0.136	51.78	0.052
+ORPO	63.66	0.421	53.21	0.063	55.43	0.143	57.17	0.180	57.80	0.152	52.08	0.049
+SimPO	63.00	0.412	52.76	0.054	56.22	0.162	57.64	0.188	**58.53**	**0.165**	**52.53**	**0.070**

Table 5: Experimental results on different preference optimization methods. The best results for each group are in **bold**. The second-best results for each group are with underline.

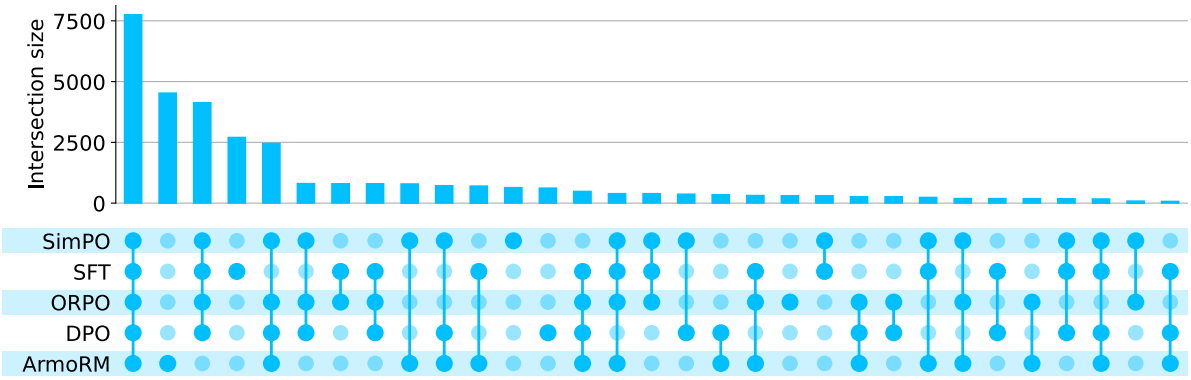
nism of preference alignment. However, the distributions exhibit notable differences: while SimPO shows an unintended reduction in likelihoods for preferred responses, DPO achieves superior separation through what appears to be a linear reorganization of the probability space. This distinctive pattern suggests DPO's particularly effective transformation of the model's internal representation space for preference modeling. The comparative performance highlights DPO's ongoing potential for preference alignment tasks and underscores the value of further investigating its underlying optimization dynamics.



(a) UniHypo (w/o Length Normalized)

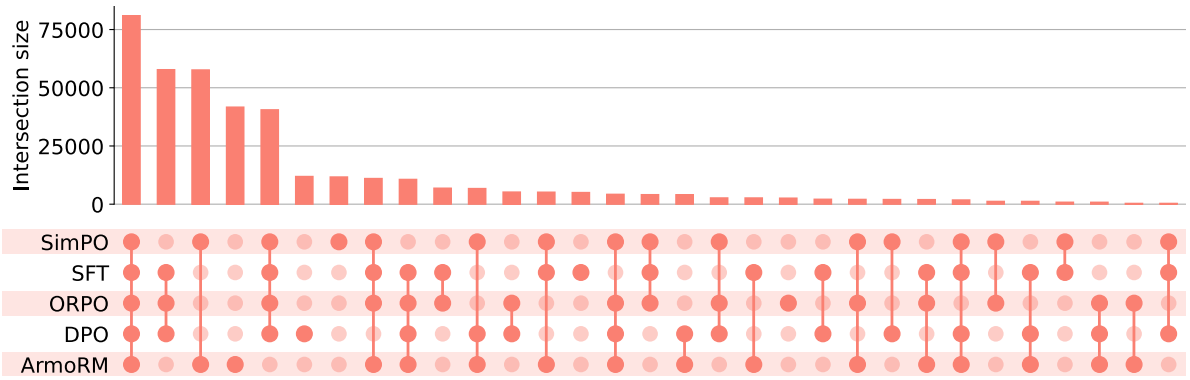


(b) HelpSteer2 (w/o Length Normalized)

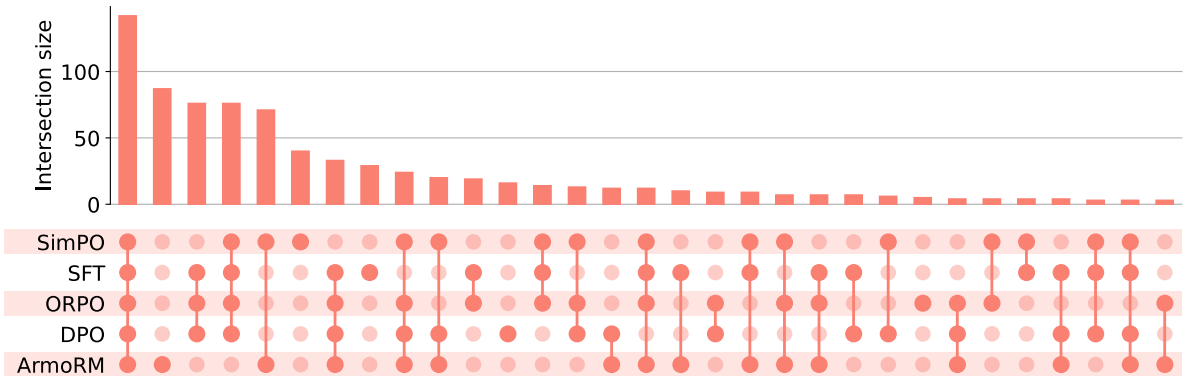


(c) UltraFeedback (w/o Length Normalized)

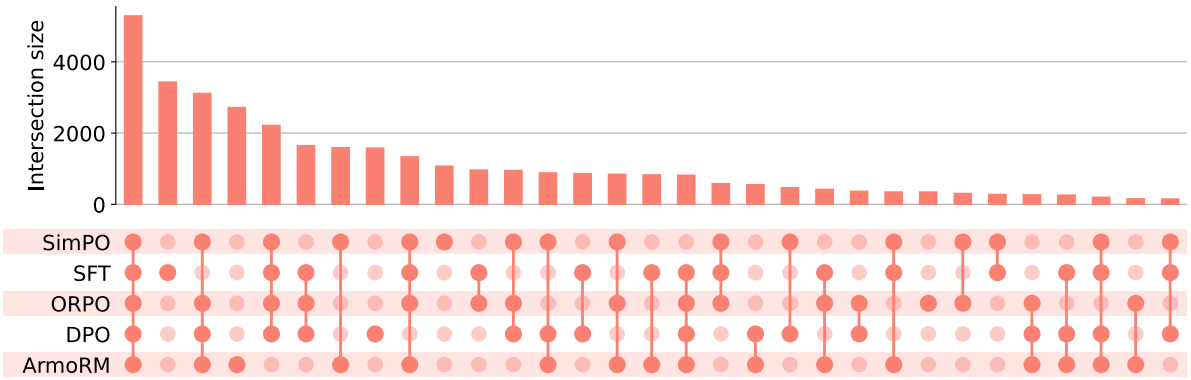
Figure 4: Upset plots of generation likelihoods without length normalization.



(a) UniHypo (w/ Length Normalized)

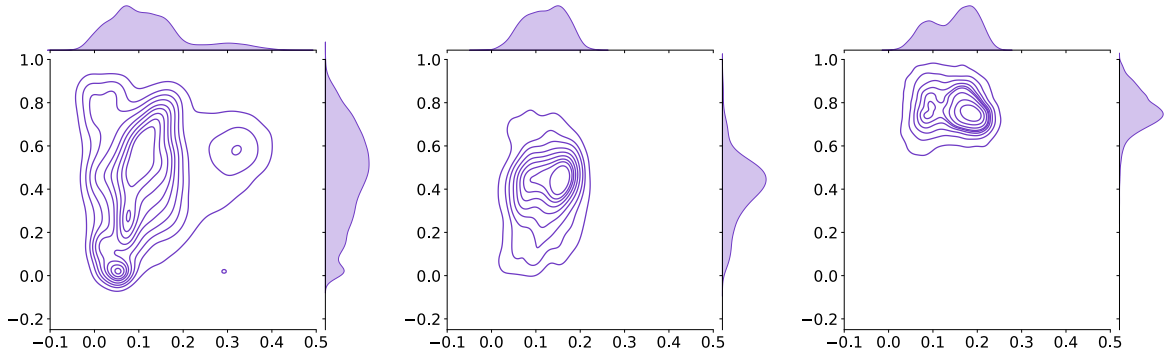


(b) HelpSteer2 (w/ Length Normalized)

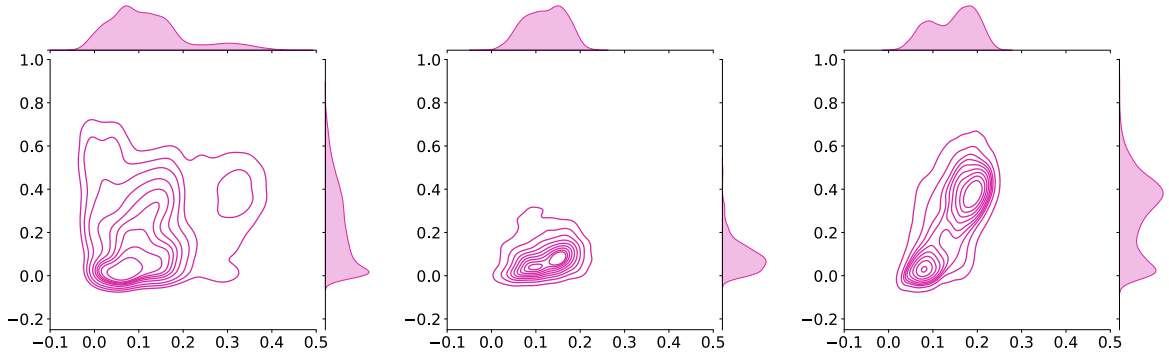


(c) UltraFeedback (w/ Length Normalized)

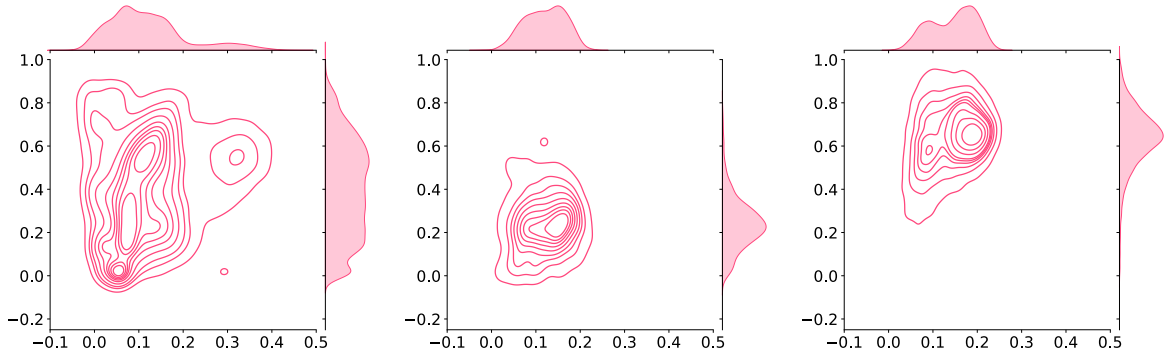
Figure 5: Upset plots of generation likelihoods with length normalization.



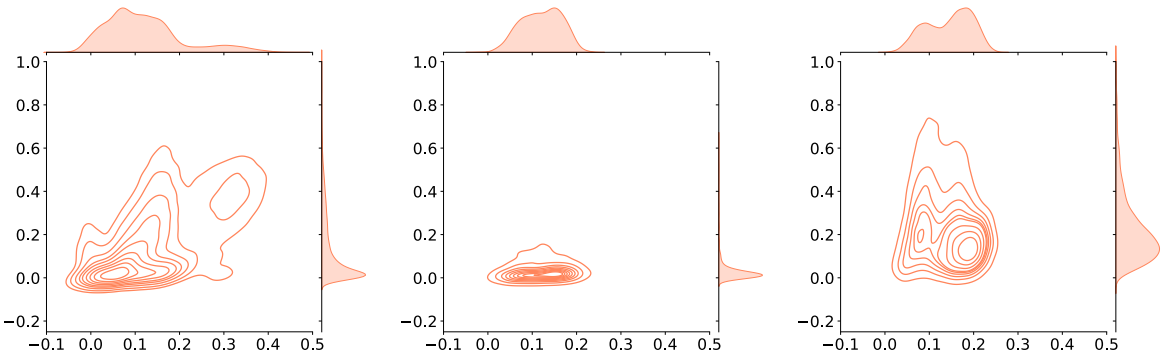
(a) UniHypo - SFT (b) HelpSteer2 - SFT (c) UltraFeedback - SFT



(d) UniHypo - DPO (e) HelpSteer2 - DPO (f) UltraFeedback - DPO



(g) UniHypo - ORPO (h) HelpSteer2 - ORPO (i) UltraFeedback - ORPO



(j) UniHypo - SimPO (k) HelpSteer2 - SimPO (l) UltraFeedback - SimPO

Figure 6: Joint plots of generation likelihoods and reward scores.