Clear Preferences Leave Traces: Reference Model-Guided Sampling for Preference Learning

Anonymous submission

Abstract

Direct Preference Optimization (DPO) has emerged as a defacto approach for aligning language models with human preferences. Recent work has shown DPO's effectiveness relies on training data quality. In particular, clear quality differences between preferred and rejected responses enhance learning performance. Current methods for identifying and obtaining such high-quality samples demand additional resources or external models. We discover that reference model probability space naturally detects high-quality training samples. Using this insight, we present a sampling strategy that achieves consistent improvements ($+0.1$ to $+0.4$) on MT-Bench while using less than half (30-50%) of the training data. We observe substantial improvements ($+0.4$ to $+0.98$) for technical tasks (coding, math, and reasoning) across multiple models and hyperparameter settings.

Introduction

Preference learning aims to align Large Language Models (LLMs) with human preferences. It is applied after pretraining and supervised fine-tuning (SFT) to teach models to generate responses that better align with human expectations while preserving knowledge acquired in earlier training stages. This approach has demonstrated practical impact in improving user experience (Bai et al. 2024), implementing safety filters (Liu, Sun, and Zheng 2024; Huang et al. 2024), and moderating content (Ma et al. 2023).

Direct Preference Optimization (DPO) (Rafailov et al. 2024) is a a supervised off-policy method that has recently emerged as a leading approach to preference learning. Unlike on-policy methods, DPO directly optimizes the policy using paired preference data, where each pair contains an aligned (or *preferred*) and misaligned (or *rejected*) response. By training the model to assign higher probabilities to preferred responses, DPO effectively shapes the model's output distribution without requiring a separate reward model. This simplicity, combined with strong empirical results, has made DPO increasingly popular for language model alignment.

Models learn better when there is a clear distinction between preferred and rejected responses, while noisy or ambiguous preferences hinder learning (Ivison et al. 2024). This distinction is quantified using *preference clarity*, measured using ground truth scalar values that indicate each response's alignment or instruction-following ability. Prior work has explored various approaches to address this challenge: modifying alignment objectives (Gao, Alon, and Metzler 2024), filtering training data (Morimura et al. 2024), and improving preference data collection (Morimura et al. 2024).

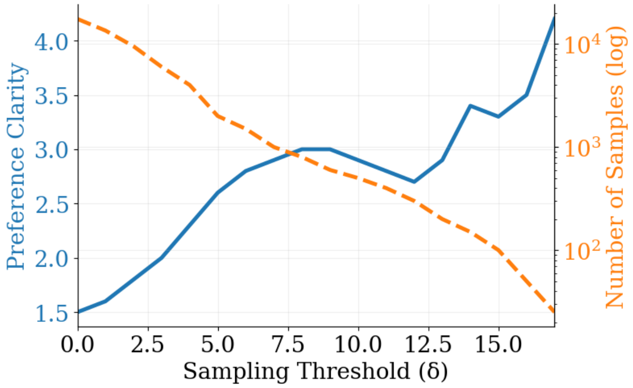

Figure 1: Relationship between sampling threshold (δ) and preference clarity for Ultrafeedback using the fine-tuned LLAMA-3 8B as the reference model. The solid blue line is the preference clarity between preferred and rejected responses calculated using the difference of Ultrafeedback's preference scores. The dashed orange line (log scale) shows the available training pairs at each threshold. **Preference pairs at a higher sampling threshold show clearer preferences, indicating that the reference model can identify high-quality preference pairs even when it incorrectly attributes the correct response.**

Creating high-quality preference pairs requires extensive annotation resources. Current datasets like Ultrafeedback (Cui et al. 2023) use GPT-4 for evaluation, but scaling this approach to new alignment datasets increases costs and creates dependencies on external models. The better way to ensure high-quality annotation is to perform manual annotation. However, this demands considerable effort — requiring multiple rounds of review, quality checks, and careful measurement of annotator agreement. These resource limitations drive the need for methods that can identify high-quality preference pairs without expensive labeling processes.

We find that the probability space of the reference model serves as a natural detector of preference clarity. When there is strong gap between the probabilities of the two responses — regardless of which it favors — the pair represents a clearer preference signal in the alignment data. This reveals an intriguing property: *the reference model can identify high quality preference pairs, even when it may not know which response is better*. Using this insight, we propose a reference model-guided sampling method that achieves better performance than the full dataset of sampling method.

We validate our findings through extensive experimentation across model architectures and hyperparameter configurations. The improvements remain consistent across these variations, suggesting our method captures fundamental preference signals rather than exploiting model-specific patterns. Using only 30-50% of the original training data, we achieve higher MT-Bench performance ($+0.1$ to $+0.4$), with particularly strong gains on technical tasks ($+0.4$ to $+0.98$ for coding, reasoning, and math). Through ablation studies, we identify key factors for effective adoption of this sampling strategy.

Methodology

Background

Direct Preference Optimization (DPO) (Rafailov et al. 2024) is a supervised off-policy method used to optimize models based on user preferences without relying on a separate reward model. Instead, DPO directly aligns the policy with preference data by leveraging a ranking-based approach.

The objective of DPO is to use the reference policy, typically a SFT model, to guide the optimization process. The preference optimization is defined using pairs of responses, where one response is preferred over the other. By focusing on these response pairs, the goal is to minimize the DPO loss function defined as:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -E_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\log\sigma\Bigg(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Bigg)\Bigg]$$

where π_θ is the policy model, π_{ref} is the reference model, σ is the sigmoid function, and (x, y_w, y_l), β is the hyperparameter controlling the deviation from the base reference policy, are preference pairs comprising a prompt x, a preferred response y_w, and a rejected response y_l, drawn from the dataset $\mathcal{D}$.

Preference Clarity It quantifies the degree of quality difference between two responses in a preference pair. Given responses r_1 and r_2, we calculate preference clarity using ground truth quality scores s_1 and s_2 assigned to each response. These scores can be obtained either through human annotation or using LLM-as-a-judge frameworks with models like GPT-4. Formally, for a preference pair (r_1, r_2), the preference clarity is measured as:

$$clarity(r_1, r_2) = |s_1 - s_2| \tag{1}$$

where s_1, s_2 are the ground truth quality scores.

Reference-model based sampling

Method. Our sampling strategy leverages normalized reference policy probabilities to identify clear preference signals. For input x with responses y_w (preferred) and y_l (rejected), we compute the difference between length-normalized reference probabilities. We sample pairs where this difference exceeds threshold δ:

$$\left|\frac{\log(\pi_{\text{ref}}(y_w|x))}{|y_w|} - \frac{\log(\pi_{\text{ref}}(y_l|x))}{|y_l|}\right| \geq \delta \tag{2}$$

where $\pi_{\text{ref}}(y|x)$ is the reference probability and $|y|$ is the sequence length. This normalization enables comparison between responses of different lengths, and focuses training on pairs with larger probability gaps. This simple approach prioritizes training examples with clearer preference signals while downweighting pairs where reference model estimates that the responses are similar.

Reference Model as Quality Detector. Our analysis reveals that the sampling strategy effectively identifies high-quality preference pairs. Using LLAMA-3-8B as the reference model, we compute probability differences between preferred and rejected responses in the Ultrafeedback dataset (Figure 1). Samples selected at higher thresholds (δ) in Equation 2 consistently demonstrate greater preference clarity. This suggests reference models naturally detect strong training examples without requiring additional annotation.

Experiments

We use a simple evaluation framework for our sampling method. First, we use an SFT model as our reference model. We compute the probabilities for all responses in our preference pairs. We then create different versions of our training dataset by sampling based on probability gaps. Each version uses a different threshold δ. We align models using DPO on both the full dataset and our sampled versions. Finally, we evaluate and compare their performances on standard benchmarks. This setup directly tests whether selecting clearer preference pairs improves model alignment.

Dataset. We use the Ultrafeedback dataset (Cui et al. 2023) for all our experiments. The dataset provides a binarized preference version containing 64k samples, where each response pair is accompanied by preference scores. These scores are generated using GPT4 in an LLM-as-a-judge framework (Zheng et al. 2023), which evaluates responses across multiple technical aspects using scalar ratings. This scoring approach has demonstrated high agreement with human annotators on technical criteria. We leverage the difference between these preference scores as a ground truth measure of preference clarity.

	Dataset		MT-Bench Performance	
Model	δ	Percentage	Score	Δ vs Full Dataset[†]
Mistral 7B	SFT	–	6.33	-0.63
	+ $\delta \geq 0$	100%	6.96	–
	+ $\delta \geq 0.5$	70%	7.13	$+0.17$
	+ $\delta \geq 1$	48%	6.97	$+0.01$
	+ $\delta \geq 2$	21%	6.99	$+0.03$
LLAMA-3-8B	SFT	–	6.55	-0.95
	+ $\delta \geq 0$	100%	7.40	–
	+ $\delta \geq 1$	57%	7.77	$+0.37$
	+ $\delta \geq 2$	31%	7.8	$+0.40$
LLAMA-3.1-8B	SFT	–	6.56	-1.21
	+ $\delta \geq 0$	100%	7.74	–
	+ $\delta \geq 1$	57%	7.88	$+0.14$
	+ $\delta \geq 2$	48%	7.85	$+0.11$

[†] Δ shows performance difference compared to using full dataset ($\delta \geq 0$).

Table 1: All the models achieve better performance ($0.1 - 0.4$) on MT-Bench with the sampled data using our reference-model sampling approach compaed to the full dataset ($\delta \geq 0$) and the SFT model.

Models and Hyperparameters We perform experiments using three different models - Mistral 7B (Jiang et al. 2023), LLAMA-3-8B (Dubey et al. 2024) and LLAMA-3.1-8B (Dubey et al. 2024). For LLAMA-3, we use the SFT checkpoint provided by SimPO (Meng, Xia, and Chen 2024) after training on Ultrachat (Ding et al. 2023) For Mistral and LLAMA-3.1, we use the Ultrachat dataset to perform SFT on the models. For table 1, we perform all our experiments with $\beta = 0.01$ for DPO. We experiment with different versions of the reference thresholds ($\delta = 0, 0.5, 1, 2$).

Evaluation We measure the performance of the model on MT-Bench (Zheng et al. 2023), a multi-turn alignment dataset that grades the answers of the policy model using GPT4 by assigning a scalar score to each response. MT-Bench categorizes their questions on 8 categories - writing, roleplay, extraction, reasoning, math, coding, knowledge I (STEM), and knowledge II (humanities/social science).

Results and Discussion

Performance

We present the results in table 1. The policy models trained using our proposed reference model-based sampling consistently outperform those trained on the full dataset. Notably, the improvements range from moderate to large ($+0.11$ to $+0.40$), with most notable gains achieved using reduced training data. For LLAMA3, we observe large improvements ($+0.40$) while using less than a third of the original dataset. Similarly, LLAMA 3.1 shows moderate gains ($+0.14$) using less than half the data. The pattern differs slightly for Mistral 7B, where the largest performance improvement occurs with 70% of the original dataset and a smaller sampling threshold.

Hyperparameters

The performance of our approach depends on both the sampling threshold (δ) and preference optimization parameters.

Sampling Threshold (δ): No single sampling threshold δ works optimally across all models. While higher thresholds identify clearer preference pairs, they also reduce the training data size, which can limit performance gains. For instance, Mistral 7B achieves optimal performance at $\delta = 0.5$, with larger thresholds showing no slight improvement over the full dataset. We note that our discrete sampling thresholds ($\delta \in \{0, 0.5, 1, 2\}$) may not exhaustively cover the optimal values for each model, suggesting potential for further optimization. However, we typically notice the largest increase in benchmark performance when 50-70 % of the original dataset is retained using the reference-based sampling.

Preference Optimization (β): In contrast, we believe a fixed $\beta = 0.01$ performs well regardless of the sampling threshold. We notice small increases ($+0.1$) in MT-Bench for LLAMA-3.1-8B when we train it using large $\beta = 0.1$. However, we do not observe an increase in performance for LLAMA-3-8B. Instead, the performance remains the same or slightly decreases (-0.1) for sampling thresholds. One possible explanation may be that a larger β imposes a stronger penalty on model deviations, potentially limiting beneficial adaptations, while a smaller β permits more flexibility, which may enhance alignment with human preferences.

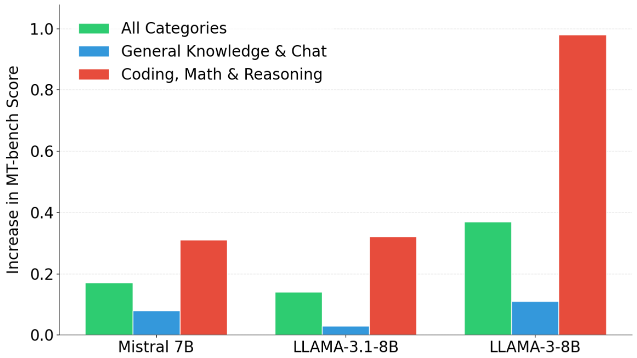

Figure 2: Performance improvements across different task categories for the best version of sampling approach , measured by increase in MT-bench scores. Technical tasks (Coding, Math & Reasoning) show substantially larger gains compared to general tasks.

Aspect-wise increases

We also analyze task-specific performance improvements on MT-Bench. Technical tasks (coding, math & reasoning) show large gains compared to general knowledge and chat tasks (writing, roleplay, extraction, knowledge). LLAMA3-8B demonstrates the most dramatic improvements, with MT-Bench scores for technical tasks increasing by nearly 1.0 points. While Mistral and LLAMA-3.1 show more modest gains, their improvements on technical tasks remain substantial (+0.4). The improvements in general knowledge and chat tasks, while consistent across models, are less pronounced (+0.2 to +0.4). This disparity might be attributed to benchmark saturation – models trained on the full dataset already achieve high scores on non-technical tasks, leaving limited room for improvement. In contrast, technical tasks present more headroom for measurable gains, suggesting our sampling method is effective at identifying and leveraging strong preference signals for technical tasks.

Limitations & Future Work

There are several promising directions to build upon our work. First, testing our approach on LLMs of varying sizes and architectures would help verify the conditions under which this property holds. Experiments with models trained on different SFT datasets could further establish the generality of our findings. While MT-Bench serves as a standard benchmark for measuring alignment and instruction-following capabilities, it has limitations in consistency, reliability and biases (Zheng et al. 2023). Alternative benchmarks like Arena-hard (Li et al. 2024) could provide additional validation, though computational costs currently restrict such evaluation. It would also be interesting to verify if this property holds for other preference learning mechanisms. From a theoretical perspective, understanding why alignment techniques learn better from samples with large probability gaps in the reference model space could provide insights into preference learning mechanisms.

Related Work

Preference Learning Methods. Preference alignment has emerged as a crucial step in developing LLMs for production environments. Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022; Christiano et al. 2017) pioneered this approach, using on-policy learning to align models with human preferences. Recent work has shifted toward off-policy methods, with Direct Preference Optimization (DPO) (Rafailov et al. 2024) gaining widespread adoption due to its simplicity and effectiveness. Several variants of DPO have been proposed, introducing additional regularization terms (Liu et al. 2023; Pal et al. 2024), new hyperparameters (Meng, Xia, and Chen 2024), or modified training objectives (Ethayarajh et al. 2024).

Quality-Focused Approaches. The challenge of obtaining high-quality preference pairs has been addressed through three main approaches. The first focuses on improving data collection, using either human experts or LLMs to curate training samples (Hu et al. 2024; Huang et al. 2023; Jiang et al. 2024). The second develops more robust training mechanisms through modified objectives (Wu et al. 2024; Chowdhury, Kini, and Natarajan 2024). These approaches have seen limited adoption compared to standard DPO. The third approach involves collecting large datasets and filtering out noisy samples (Morimura et al. 2024; Kim et al. 2024). While data collection improvements require manual oversight, and robust training methods have shown limited practical success, the filtering approach offers a promising direction – especially for collecting large scale data and then narrowing down the high quality samples, preventing reiteration and expensive use of resources.

Data Quality in Alignment. Recent work has highlighted how preference data quality impacts alignment success (Ivison et al. 2024). Models demonstrate enhanced learning from clearly distinguished preference pairs, motivating research into methods for identifying and generating high-quality training data. For instance, FilterDPO (Morimura et al. 2024) proposes an on-policy approach that selects training samples by comparing policy-generated responses with preferred responses based on reward differences. While effective, such methods require additional computation or reward modeling. Our work presents a complementary approach that identifies strong preference pairs using only reference model probabilities, eliminating the need for additional resources or inference steps.

References

Bai, Z.; Wu, N.; Cai, F.; Zhu, X.; and Xiong, Y. 2024. Aligning Large Language Model with Direct Multi-Preference Optimization for Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 76–86.

Chowdhury, S. R.; Kini, A.; and Natarajan, N. 2024. Provably robust dpo: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*.

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

Cui, G.; Yuan, L.; Ding, N.; Yao, G.; Zhu, W.; Ni, Y.; Xie, G.; Liu, Z.; and Sun, M. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.

Ding, N.; Chen, Y.; Xu, B.; Qin, Y.; Zheng, Z.; Hu, S.; Liu, Z.; Sun, M.; and Zhou, B. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Gao, Y.; Alon, D.; and Metzler, D. 2024. Impact of preference noise on the alignment performance of generative language models. *arXiv preprint arXiv:2404.09824*.

Hu, Y.; Li, Q.; Ouyang, S.; Chen, G.; Chen, K.; Mei, L.; Ye, X.; Zhang, F.; and Liu, Y. 2024. Towards comprehensive preference data collection for reward modeling. *arXiv preprint arXiv:2406.16486*.

Huang, S.; Zhao, J.; Li, Y.; and Wang, L. 2023. Learning preference model for llms via automatic preference data generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9187–9199.

Huang, X.; Li, S.; Dobriban, E.; Bastani, O.; Hassani, H.; and Ding, D. 2024. One-Shot Safety Alignment for Large Language Models via Optimal Dualization. *arXiv preprint arXiv:2405.19544*.

Ivison, H.; Wang, Y.; Liu, J.; Wu, Z.; Pyatkin, V.; Lambert, N.; Smith, N. A.; Choi, Y.; and Hajishirzi, H. 2024. Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback. *arXiv preprint arXiv:2406.09279*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, R.; Chen, K.; Bai, X.; He, Z.; Li, J.; Yang, M.; Zhao, T.; Nie, L.; and Zhang, M. 2024. A Survey on Human Preference Learning for Large Language Models. *arXiv preprint arXiv:2406.11191*.

Kim, D.; Lee, K.; Shin, J.; and Kim, J. 2024. Aligning Large Language Models with Self-generated Preference Data. *arXiv preprint arXiv:2406.04412*.

Li, T.; Chiang, W.-L.; Frick, E.; Dunlap, L.; Wu, T.; Zhu, B.; Gonzalez, J. E.; and Stoica, I. 2024. From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline. *arXiv preprint arXiv:2406.11939*.

Liu, T.; Zhao, Y.; Joshi, R.; Khalman, M.; Saleh, M.; Liu, P. J.; and Liu, J. 2023. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*.

Liu, Z.; Sun, X.; and Zheng, Z. 2024. Enhancing llm safety via constrained direct preference optimization. *arXiv preprint arXiv:2403.02475*.

Ma, H.; Zhang, C.; Fu, H.; Zhao, P.; and Wu, B. 2023. Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning. *arXiv preprint arXiv:2310.03400*.

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

Morimura, T.; Sakamoto, M.; Jinnai, Y.; Abe, K.; and Air, K. 2024. Filtered direct preference optimization. *arXiv preprint arXiv:2404.13846*.

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

Pal, A.; Karkhanis, D.; Dooley, S.; Roberts, M.; Naidu, S.; and White, C. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Wu, J.; Wang, X.; Yang, Z.; Wu, J.; Gao, J.; Ding, B.; Wang, X.; Jin, R.; and He, X. 2024.

$$\backslash$$

alpha $-DPO : AdaptiveRewardMarginisWhatDirectPreference$

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.