ORPO: Monolithic Preference Optimization without Reference Model

Jiwoo Hong, Noah Lee, James Thorne
KAIST AI

Year of Publication :- 2024
Number of Citations :- 82

Introduction

- Recent preference alignment techniques for language models have demonstrated promising results [1-3].
- Supervised fine tuning (SFT) remains important for achieving convergence in these methods.
- This paper emphasizes that a minor penalty for disfavoured style is enough for preference alignment and no separate preference training step is needed.
- This paper introduce a straightforward reference model-free monolithic odds ratio preference optimization algorithm (ORPO), eliminating the need for an additional preference alignment phase.

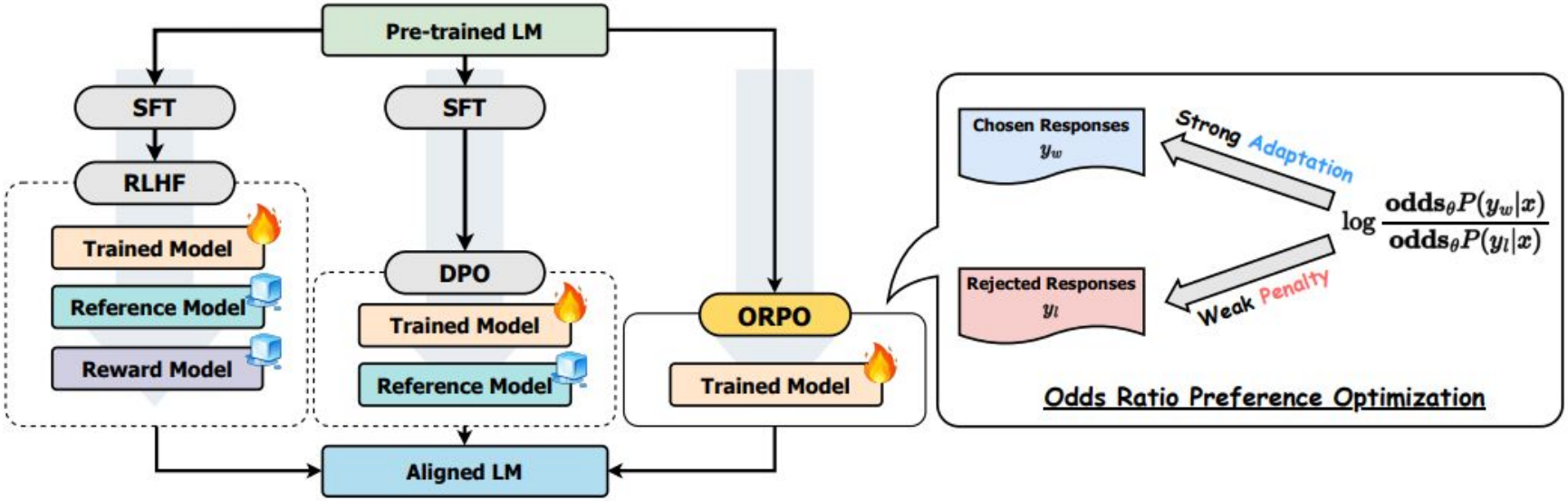
[1] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Fine tuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.
[2] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.
[3] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Related Works

- Alignment with Reinforcement Learning [2,4,5] .
  - Bradley-Terry model [9] to estimate the probability of a pairwise competition between two independently evaluated instances.
  - Reinforcement algorithm such as proximal policy optimization [10].
- Alignment without Reward Model.
  - Direct Preference Optimization [3].
  - Identity preference Optimization [6].
  - Preference alignment without pairwise preference dataset [7,8].
- Alignment with Supervised Fine tuning.
  - Supervised fine tuning with filtered datasets that align with human preference [11,12].

[2] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.
[3] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.
[4] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... & Christiano, P. F. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33, 3008-3021.
[5] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.
[6] Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., & Calandriello, D. (2024, April). A general theoretical paradigm to understand learning from human preferences. In International Conference on Artificial Intelligence and Statistics (pp. 4447-4455). PMLR.
[7] Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., & Kiela, D. Model Alignment as Prospect Theoretic Optimization. In Forty-first International Conference on Machine Learning.
[8] Cai, T., Song, X., Jiang, J., Teng, F., Gu, J., & Zhang, G. (2023). ULMA: Unified Language Model Alignment with Demonstration and Point-wise Human Preference. arXiv preprint arXiv:2312.02554.
[9]] Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika, 39(3/4), 324-345.
[10] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
[11] Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., ... & Levy, O. (2024). Lima: Less is more for alignment. Advances in Neural Information Processing Systems, 36.
[12] Li, X., Yu, P., Zhou, C., Schick, T., Levy, O., Zettlemoyer, L., ... & Lewis, M. (2023). Self-alignment with instruction backtranslation. arXiv preprint arXiv:2308.06259.

Odds Ratio Preference Optimization



Odds Ratio Preference Optimization

- Adding an odds ratio based penalty to conventional SFT loss.
- This penalty helps to differentiate between favoured and disfavoured generation styles.
- The average log likelihood of generating a sequence y of length m tokens is given below.

$$\log P_\theta(y|x) = \frac{1}{m}\sum_{t=1}^{m} \log P_\theta(y_t|x, y_{<t})$$

Odds Ratio Preference Optimization

- Odds value $$\mathbf{odds}_\theta(y|x) = \frac{P_\theta(y|x)}{1 - P_\theta(y|x)}$$

- If odds value = k, it means that is k times more likely for the model to generate the output sequence y than not generating it.
- Odds ratio

$$\mathbf{OR}_\theta(y_w, y_l) = \frac{\mathbf{odds}_\theta(y_w|x)}{\mathbf{odds}_\theta(y_l|x)}$$

Odds Ratio Preference Optimization

- Odds ratio indicates that how much likely it is for the model to generate y_w over y_l given input x.
- Objective function of ORPO is given as,

$$\mathcal{L}(d;\theta) = \mathcal{L}_{SFT}\left(x, y_w; \theta\right) + \lambda \mathcal{L}_{OR}(d;\theta)$$

Where,

$$\mathcal{L}_{OR}\left(d;\theta\right) = -\log \sigma \left(\log \frac{\mathbf{odds}_\theta(y_w|x)}{\mathbf{odds}_\theta(y_l|x)} \right)$$

Experimental Settings - Training configurations

- Phi-2 (2.7 B) [13] , Llama-2 (7 B) [14] and Mistral (7 B) [15] which are on Binarized Ultra feedback [17].
- A series of OPT models scaling 125 M to 1.3 B parameters are fine tuned on Anthropic's HH-RLHF [16] and Binarized Ultra feedback. These are fine tuned with PPO, DPO, supervised fine tuning and ORPO to compare the performance.

[13] Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C. C. T., ... & Gopi, S. (2023). Phi-2: The surprising power of small language models. Microsoft Research Blog, 1(3), 3.
[14] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
[15] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.
[16] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.
[17] Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., ... & Wolf, T. (2023). Zephyr: Direct distillation of lm alignment. arXiv preprint arXiv:2310.16944.

Leaderboard Evaluation

- Models were using the $\text{AlpacaEval}_{1.0}$ and $\text{AlpacaEval}_{2.0}$ [18] benchmarks comparing the ORPO model to other models reported in the official leaderboard.
- Similarly models are evaluated with the MT bench to report the results and the scores of the same models in official leaderboard.

Experimental Results - Single Turn Instruction Following

- ORPO improved Phi-2 (2.7 B) exceeded the performance of Llama-2 (7B) chat instruction model (ORPO lamda value = 0.25).
- Using the Ultrafeedback data and ORPO in Llaama-2 (7B) with lambda of 0.2 with Llama-2 resulted in even higher AlpacaEval scores than the chat versions of both the 7B and 13B scale trained with RLHF.
- Mistral-ORPO-alpha and Mistral-ORPO-Beta outperformed the Zephyr series models which are formed by doing SFT on Ultrachat and DPO on full ultra feedback.

Experimental Results - Single Turn Instruction Following

Model Name	Size	$\text{AlpacaEval}_{1.0}$	$\text{AlpacaEval}_{2.0}$
Phi-2 + SFT	2.7B	48.37% (1.77)	0.11% (0.06)
Phi-2 + SFT + DPO	2.7B	50.63% (1.77)	0.78% (0.22)
Phi-2 + ORPO *(Ours)*	2.7B	**71.80% (1.59)**	**6.35% (0.74)**
Llama-2 Chat *	7B	71.34% (1.59)	4.96% (0.67)
Llama-2 Chat *	13B	81.09% (1.38)	7.70% (0.83)
Llama-2 + ORPO *(Ours)*	7B	**81.26% (1.37)**	**9.44% (0.85)**
Zephyr (α) *	7B	85.76% (1.23)	8.35% (0.87)
Zephyr (β) *	7B	90.60% (1.03)	10.99% (0.96)
Mistral-ORPO-α *(Ours)*	7B	87.92% (1.14)	11.33% (0.97)
Mistral-ORPO-β *(Ours)*	7B	**91.41% (1.15)**	**12.20% (0.98)**

Experimental Results - Multi-turn Instruction Following

- The best models Mistral ORPO are assessed for multi turn instruction following skills with deterministic answers (in Math).
- This is tested through MT bench.
- In results, it is shown that Mistral-ORPO series exceeds larger instruction following chat models, especially Llama-2-chat (70B).

Experimental Results - Multi-turn Instruction Following

MT-Bench	1st Turn	2nd Turn	Average
Llama-2-7B Chat	6.41	6.13	6.27
Llama-2-13B Chat	7.06	6.24	6.65
Llama-2-70B Chat	6.99	6.73	6.86
Mistral-ORPO-α	7.49	6.96	7.23
Zephyr-β	**7.68**	6.98	**7.33**
Mistral-ORPO-β	7.64	**7.00**	7.32

Experimental Results - Case study with smaller models

- Win rates of ORPO was assessed over other preference alignment techniques using different scales of OPT modules.

ORPO vs	SFT	+DPO	+PPO
OPT-125M	84.0 (0.62)	41.7 (0.77)	66.1 (0.26)
OPT-350M	82.7 (0.56)	49.4 (0.54)	79.4 (0.29)
OPT-1.3B	78.0 (0.16)	70.9 (0.52)	65.9 (0.33)

Analysis - Computational Efficiency

- Computational costs for training DPO and ORPO on Mistral (7B) for 1 epoch on Ultra Feedback dataset was calculated.
- It was done using AdamW with DeepSpeed ZeRO 3.
- Also, SFT training time of DPO was excluded for comparison.

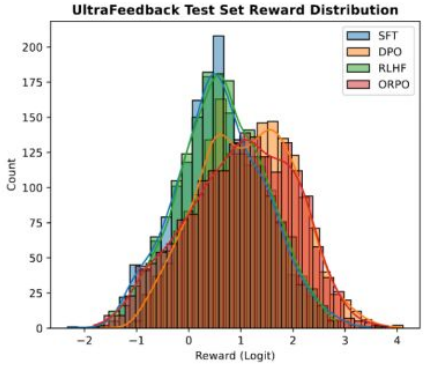
	DPO	ORPO
Training Time (hours) (↓)	12.6	**5.5**
Max Batch (↑)	1	**4**

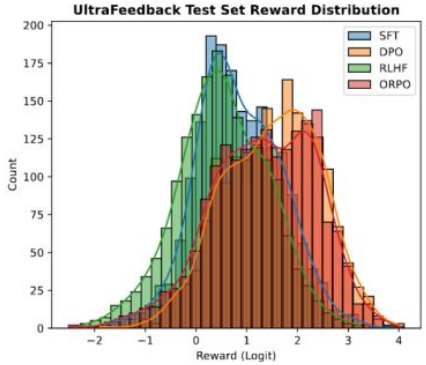
Analysis - Lexical Diversity

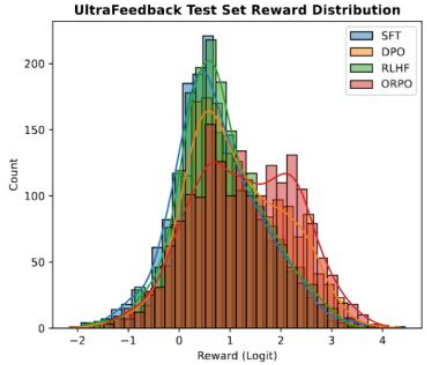
- Per Input Diversity.
  - For input, a number of sample outputs are generated and average cosine similarity between the outputs are calculated.
- Across input diversity.
  - 8 samples are generated per input, first item of each input are taken and cosine similarity is taken and averaged.

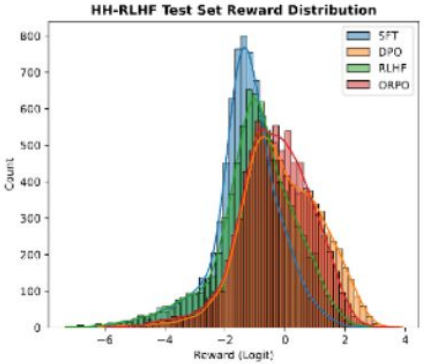
	Per Input↓	Across Input↓
Phi-2 + SFT + DPO	**0.8012**	0.6019
Phi-2 + ORPO	0.8909	**0.5173**
Llama-2 + SFT + DPO	**0.8889**	0.5658
Llama-2 + ORPO	0.9008	**0.5091**

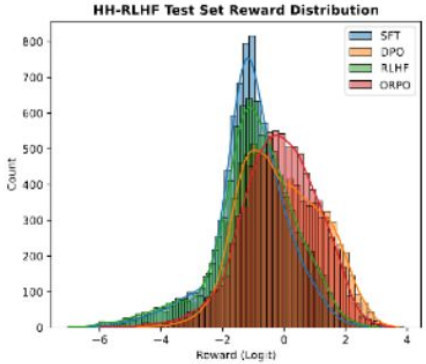
Analysis - Minimizing Reward Distribution

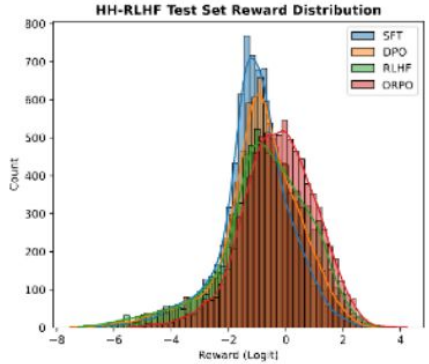




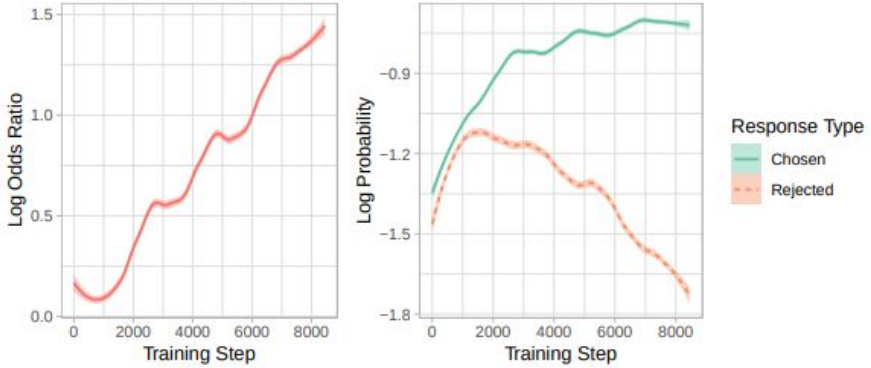








Analysis - Minimizing Odds ratio penalty



Conclusion

- This paper introduced a reference free monolithic preference alignment technique known as Odds Ratio Preference Optimization.
- The method is shown to provide better results when compared to the existing techniques for preference alignment.

Limitations and Future Work

- Comprehensive analysis was done for preference alignment techniques including DPO and RLHF, but this analysis was not done on a comprehensive range of preference alignment techniques.
- Fine tuning datasets will be expanded to different domains and qualities.
- Analyzing the internal impact of the method on the pre-trained language model.

References

1. Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Fine tuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.
2. Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.
3. Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.
4. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... & Christiano, P. F. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33, 3008-3021.
5. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.
6. Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., & Calandriello, D. (2024, April). A general theoretical paradigm to understand learning from human preferences. In International Conference on Artificial Intelligence and Statistics (pp. 4447-4455). PMLR.
7. Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., & Kiela, D. Model Alignment as Prospect Theoretic Optimization. In Forty-first International Conference on Machine Learning.
8. Cai, T., Song, X., Jiang, J., Teng, F., Gu, J., & Zhang, G. (2023). ULMA: Unified Language Model Alignment with Demonstration and Point-wise Human Preference. arXiv preprint arXiv:2312.02554.
9.
10. Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika, 39(3/4), 324-345.
11. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
12. Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., ... & Levy, O. (2024). Lima: Less is more for alignment. Advances in Neural Information Processing Systems, 36.
13. Li, X., Yu, P., Zhou, C., Schick, T., Levy, O., Zettlemoyer, L., ... & Lewis, M. (2023). Self-alignment with instruction backtranslation. arXiv preprint arXiv:2308.06259.
14. Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C. C. T., ... & Gopi, S. (2023). Phi-2: The surprising power of small language models. Microsoft Research Blog, 1(3), 3.
15. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
16. Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.
17. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.
18. Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., ... & Wolf, T. (2023). Zephyr: Direct distillation of lm alignment. arXiv preprint arXiv:2310.16944.
19. Li, X., Zhang, T., Dubois, Y., Taori, R., Ishaan Gulrajani, C. G., Liang, P., & Hashimoto, T. B. Alpacaeval: An Automatic Evaluator of Instruction-Following Models. 2023. URL https://github. com/tatsu-lab/alpaca_eval.