Literature Review: ORPO: Monolithic Preference Optimization without Reference Model

Luca Mouchel – 324748 – luca.mouchel@epfl.ch
Group: REAL

1 Summary

This paper introduces a novel approach called Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024) for preference alignment in language models. The authors highlight the importance of supervised fine-tuning (SFT) in achieving successful convergence and emphasize the need to penalize undesired generation styles during SFT. In NLP, especially in the alignment phase, SFT is critical. It's role is to fine-tune language models to generate sequences aligning with the training data.

The key contribution of this paper is eliminating the need for a reference model, contrary to other largely used preference alignement methods, including PPO (Schulman et al., 2017) and DPO (Rafailov et al., 2023). DPO was already a major step forward, allowing to outperform PPO on downstream tasks, all the while skipping the model rewarding phase.

The authors propose ORPO as a monolithic alignment method that dynamically penalizes dispreferred responses without the need for a reference model. The objective function of ORPO combines the SFT loss with a relative ratio loss, which maximizes the odds ratio between favored and disfavored responses. The effectiveness of ORPO is demonstrated through empirical evaluation, showing that it outperforms state-of-the-art language models in various downstream tasks, achieving significant improvements in AlpacaEval2.0 (Li et al., 2023), and MT-Bench (Zheng et al., 2023). The paper concludes with a discussion on the theoretical analysis, computational efficiency, and future work for ORPO.

The experimental settings section describes the training configurations and datasets used in the study. The authors train a series of models with different algorithms, including SFT, PPO, DPO, and ORPO (Schulman et al., 2017; Rafailov et al., 2023; Hong et al., 2024). They evaluate the models on two datasets and compare their performance with other models reported in official leaderboards and show the performance of ORPO increases significantly compared to SFT + DPO, as shown in Table 1 with the Phi-2 model. The authors also provide details on the reward models used for evaluation. The results demonstrate the effectiveness of ORPO in achieving high performance in various benchmarks, surpassing state-of-the-art models in different tasks. The paper concludes with a discussion on the theoretical, empirical, and computational justification of ORPO and the release of code and model checkpoints for further research and development.

2 Strengths

One of the strengths of this paper is the introduction of a novel approach called Odds Ratio Preference Optimization (ORPO) for preference alignment in language models. ORPO eliminates the need for a reference model and offers a monolithic alignment method that dynamically penalizes disfavored responses (Hong et al., 2024). This approach is innovative and addresses the limitations of existing methods that require multiple stages and reference models. For example, PPO requires a reference model (Schulman et al., 2017) (by running supervised fine-tuning), a reward model, and the optimization phase. More recently, DPO (Rafailov et al., 2023) was published and presented a new alignment method skipping the reward modeling phase and needing only a reference model and running the optimization phase. ORPO presents the next step up, showing we can align models from preference data, without requiring a reference model.

By incorporating an odds ratio-based penalty into the negative log-likelihood loss, ORPO effectively differentiates between favored and disfavored generation styles during supervised fine-tuning (SFT). This not only improves the alignment procedure but also ensures the preservation of domain adaptation during the training process. The paper also provides a comprehensive evaluation and comparison of ORPO with other alignment methods, demonstrating its effectiveness in achieving high performance in different tasks. Additionally, the paper includes a thorough literature review of related works, theoretical justifications, and computational justifications for the use of odds ratio in preference alignment, further strengthening the credibility and robustness of the proposed method.

3 Weaknesses

As mentioned in the limitation section of the paper, because of the small amount of alignement methods commonly used, this represents a weakness since there are not many benchmarks to compare ORPO's performance with (Hong et al., 2024). They also mention scaling the fine-tuning and optimization phase to larger models (since they only worked with $\leq$7B models). They additionally mention the necessity to expand the fine-tuning datasets into diverse domains and qualities to verify the generalizability of the method in various NLP downstream tasks. Additionally, the paper acknowledges the need to study the internal impact of the method on the pre-trained language model, expanding the understanding of the preference alignment procedure to not only the supervised fine-tuning stage but also consecutive preference alignment algorithms. These limitations indicate areas for future work and improvement in the research.

4 Conclusion

Overall, this paper presents some promising results and carves the path towards more simplistic yet increasingly more performative alignement methods for preference optimization. By removing the need for a reference model, preference optimization becomes more lightweight and easier to implement.

References

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.