Literature Review: Disentangling Length from Quality in Direct Preference Optimization

André E. Santo | 376762 | andre.loureiroespiritosanto@epfl.ch
My Nice Long Penguin

1 Summary

Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) are two well-known techniques to finetune a pre-trained model. This paper (Park et al., 2024) presents two main contributions: The first in-depth study of the length bias in DPO, showing that this bias applies not only to RLHF but also to DPO, and a regularization method that promises to alleviate this problem. As pointed out by the authors, this problem is relevant not only from the theoretical standpoint but also from the fact that DPO is the main algorithm used by open-source models as it is easier to implement and less expensive. Hence, this regularization method promises to increase the performance of these models to match that of proprietary models.

They begin by explaining how the RLHF pipeline works and provide its mathematical foundations. Afterward, they show how the DPO objective is derived from the RLHF. Thereafter, the researchers compare the mean length of answers given by the original DPO algorithm and the mean length of preferred and rejected answers, evaluated by humans. They state that, indeed, DPO produces out-of-distribution lengths, with a high tendency to be bigger.

To assess the performance of the regularization method and the aforementioned relation, the authors use 2 tasks: Summarization (TL;DR dataset (Stiennon et al., 2022)) and Dialogue (HH dataset (Bai et al., 2022)). However, the researchers could also have used a dataset based on a more open-ended task (eg.: storytelling) to assess how the method performs under instances where verbosity might be a proxy for quality and not just a bias.

This regularization technique obtains considerable results, achieving a 20% increase in win rates when controlling for length, even when the judge, GPT-4, is biased towards verbosity.

Subsequently, the researchers study whether length is a proxy for KL-Divergence. While the results did not show a clear correlation between both, it is hypothesized that this may be caused by other factors driving human preference. Nonetheless, a clearer and more in-depth explanation of this phenomenon could have been done.

The authors also present the evolution of *KL-divergence*, *Mean length*, and *Mean GPT-4 Win Rates* throughout an epoch. It is visible that DPO without the regularization term (henceforth called classic DPO) grows in length much faster than the regularized one. Furthermore, classic DPO starts with a bigger win rate (due to GPT-4 being biased towards verbosity) but eventually gets surpassed by the regularized version, while keeping a smaller length, indicative that it is exploring other qualities that the first version could not. Finally, KL-Divergence increases faster without length regularization.

This work is a considerable contribution to the field. However, it would benefit from using other judges that are not as biased as GPT-4 to be able to make a comparison.

The authors do not shy away from their work's limitations, stating that only DPO and verbosity bias were studied, so no other algorithms or biases are addressed in their work.

2 Strengths

The goal of the paper is clearly explained. The authors explain how the RLHF pipeline works, the RL objective function, and how the DPO objective function is derived from RLHF, explaining that it assumes access to the optimal reward and policy. Furthermore, the paper explains that DPO is more resource-efficient, less expensive to train and easier to implement, which makes it a popular choice for open-source models. The authors claim that "*9 out of the top 10 models on the HuggingFace Open LLM Leaderboard use DPO as part of*

their training pipeline.". This introduction frames the importance of studying how length bias affects DPO and how to solve it.

This study offers two main contributions to the field of Modern Natural Language Processing (MNLP), particularly in our understanding of DPO limitations and solutions for those limitations. Firstly, it provides proof that DPO, like RLHF, suffers from length exploitation, linking it to out-of-distribution bootstrapping.

Afterward, the study proposes a method to tackle this issue, presenting a regularization factor added to the DPO objective function. This factor has a hyperparameter α which is inversely proportional to the average length of the model output. This is a considerable contribution as it allows open-source models using DPO to match the performance of proprietary models using RLHF, as the authors claim.

The authors then use their methodology to evaluate their proposal's effectiveness experimentally. They used 256 samples from each dataset. Given the level of detail in the description of the experiments, it should be fairly straightforward to reproduce the results achieved by the authors. They state that they use the original DPO codebase, providing the link to the online repository. The datasets were identified, as well as the training hyperparameters.

The plots presented to support their work are comprehensive and interpretable. They are comprehensively explained and analyzed as well. The results obtained in most of their experiments are concordant with the expectations.

All in all, the analysis of the obtained results is correct and corresponds to the authors' claims.

3 Weaknesses

The first weakness worth noticing is the number of lapses in the text. Sometimes, the authors point out to values that are not concordant with their setting (eg.: writing $\beta = 0.01$ instead of $\beta = 0.1$ in Section 4.3). While these are understandable lapses, they might make interpretation more difficult for the reader. Additionally, the figures could be better positioned in the paper, as they are often referenced several pages after they are presented.

Upon experimentally testing their regularization parameter, the authors use 2 datasets that depict Summarization and Dialogue tasks. However, their results would be better established if they were also assessed in tasks where verbosity is more of a requirement than a bias, such as storytelling. This would allow to assess the generalization ability of the parameter and understand whether the model still shows verbosity where it is needed and where verbosity is a proxy for quality. However, one can also interpret the lack of this type of task as an indication that the method should not be used to finetune a model that actually should yield verbosity. Testing on such a task would clear out this doubt.

The baseline used to compare their results is classic DPO and a Supervised Fine Tuning (SFT) model. However, since the authors establish comparisons with RLHF, a model trained with RLHF could be used to directly assess whether the claims about open-source models matching RLHF performance are indeed true.

To obtain the win rates for each model, the authors used GPT-4 as a judge. Indeed, given that the way the win rates are obtained and what prompt is used may have an impact on the results, the researchers could have disclosed more information on how this was performed, as it would have strengthened their results. Moreover, the only model used as a judge is GPT-4. As this model has a known verbosity bias, the model chosen by the authors is adversarial. However, other models could have been used to understand how these rates fluctuate when another model (especially one with less verbosity bias) is used as a judge.

In Section 4.4, the authors reason about whether length is a proxy for KL-Divergence. However, the results did not meet the expectations. As the authors state, "*For both HH and TL;DR, length-regularized models trained with β = 0.05 and β = 0.01 match the average length of train runs with β = 0.5 (Fig. 3). At the same time, these runs have statistically significant higher KL divergences and win rates*" The only explanation provided is "*We hypothesize that this indicates the existence of different factors driving human preference, with length being only a partial one.*". The work would benefit from a more detailed explanation, as they fail to completely acknowledge and explain the limitations of their claim that the exploitation of the length bias by DPO is linked to *out-of-distribution bootstrapping*.

Finally, the researchers do not present any ethical considerations. This would have added value and completeness to their study.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

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

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback.