Exploring Coherence-Verbosity Trade-off in Multi-objective DPO Training for LLMs

An Doan, Felicity Huang, Linda Liu

CS224R, Department of Computer Science, Stanford University

Stanford
Computer Science

Project Overview

- **Motivation:** Large Language Models (LLMs) have been widely embraced by users worldwide. Preference-based RL usually focus on a single reward signal, but alignment with other values like correctness and helpfulness can also be important.
- **Objective:** We investigated ways to improve LLM capabilities in instruction-following and explored trade offs between multiple-objective reward signals, specifically **coherence** and **verbosity**.
- **Setup**: We used Supervised Fine-tuning (SFT) and Direct Preference Optimization (DPO) to improve upon a Qwen2.5 0.5B base model. We then trained the fine-tuned model using multi-objective DPO, optimizing for both coherence and verbosity.
- **Findings:** DPO improved the win-rate significantly by 25%. Multi-objective DPO on new dataset improved performance by an additional 20%. The dataset that prioritized verbosity over coherence resulted in higher DPO performance, though all win rates were around 80%.

Datasets & Metrics

Training: We used three different datasets in our SFT, DPO and multi-objective DPO components.

- **SmolTalk (SFT):** More than 1.1 million prompt-answer pairs taken from high-quality chat responses from GPT-4o
- **UltraFeedback (DPO):** More than 61,100 chosen-reject answer pairs to user prompts
- **HelpSteer (Multi-objective DPO):** Over 35,000 prompt-response pairs with human-annotated scores for helpfulness, correctness, coherence, complexity, and verbosity

Evaluation: We used a Llama 3.1 Nemotron 70B Reward Model to score our model's responses on 2,000 held-out UltraFeedback prompts and calculate a "win rate," or when our model performs better than some baseline model.

Method	Win-rate baseline model
SFT	Qwen 2.5 0.5B Instruct
DPO	Qwen 2.5 0.5B + SFT
Multi-objective DPO	Qwen 2.5 0.5B + SFT

Methods & Experiments

Supervised Fine-tuning (SFT)

To perform SFT, we used a next-token prediction objective (below) to maximize the log likelihood of our favored responses, or the expert distribution, given the prompt. We used SmolTalk as the expert distribution and completed 1 epoch of training.

$$\max_{\theta} \mathbb{E}_{x,y\in D} \sum_{t=1}^{|y|} \log \pi_\theta(y_t|x, y_{<t})$$

Direct Preference Optimization (DPO)

DPO has been demonstrated as a lightweight, state-of-the-art method to align LMs with human preferences.[4] To perform DPO, we used our SFT-improved Qwen model as a reference model and used an objective (below) that maximizes the relative log likelihood of the model producing human "chosen" responses over "rejected" responses, given the prompt. We completed one epoch of training using a beta value of 0.1.

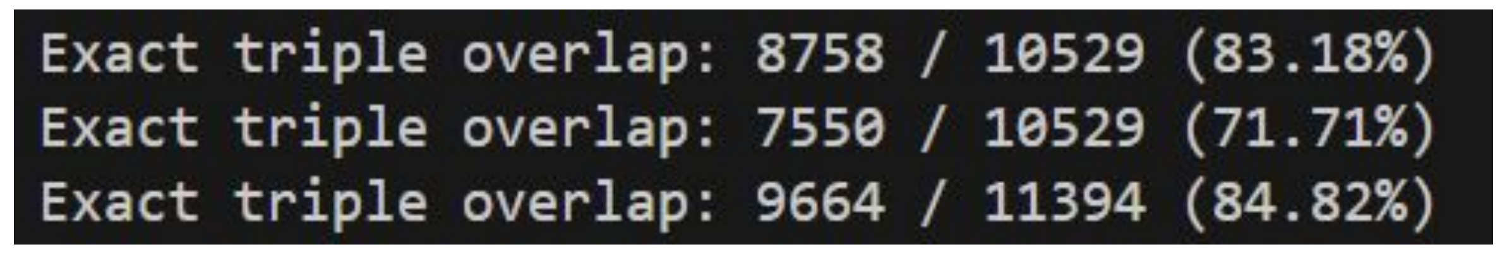
$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

Multi-objective DPO

Scholars have proposed the benefit of using multi-objective benchmarks and data in training.[5] We used the following weighted sum scalarization reward to explore the tradeoff between coherence and verbositv.

$$r_{total} = \lambda \times r_{coherence} + (1 - \lambda) \times r_{verbosity}$$

We explored 3 coherence-verbosity ratios, A: 30%/70%, B: 50%/50%, C: 70%/30%. For each, we generated a preference dataset using the HelpSteer dataset and used DPO to train our model on the new dataset. The overlap of exact prompt, chosen, rejected triples between the created datasets A-B, A-C, B-C were as follows:

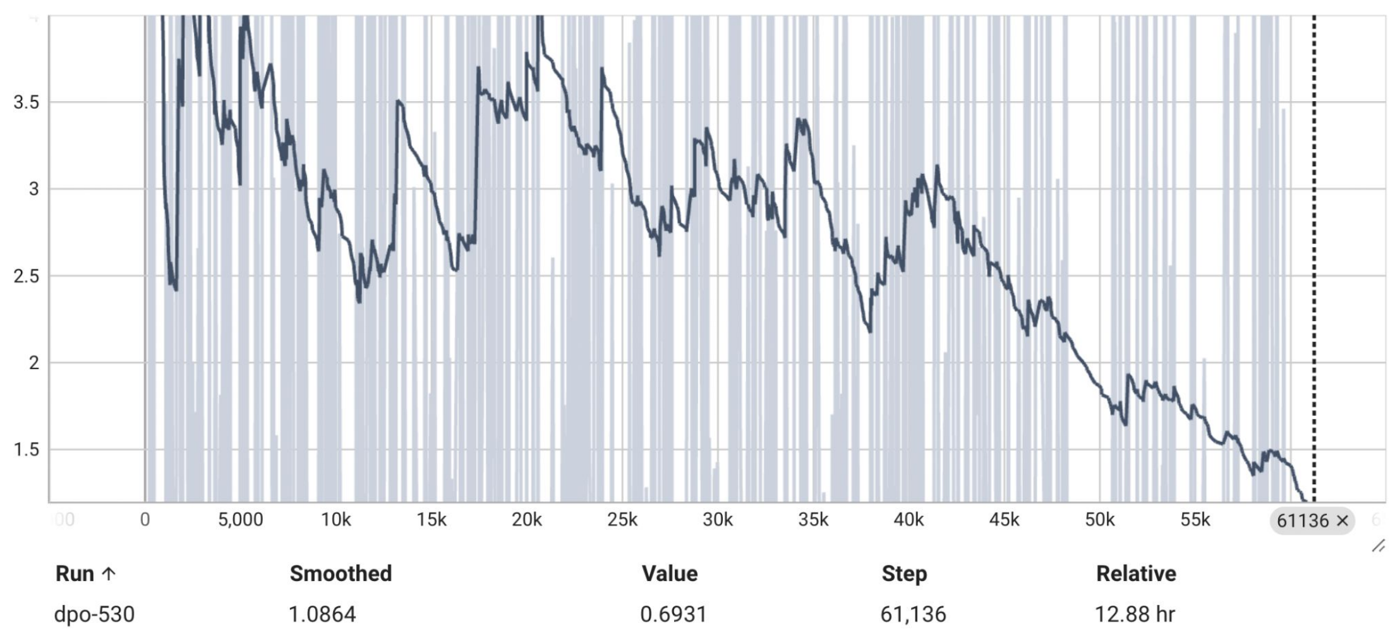


```
Exact triple overlap: 8758 / 10529 (83.18%)
Exact triple overlap: 7550 / 10529 (71.71%)
Exact triple overlap: 9664 / 11394 (84.82%)
```

Experiment Details

For each method, we completed one epoch of training using a batch size of 1, beta value of 0.1, and learning rate of 1e-6. From graph shown on the right, we can see that the loss of training LLMs on preference datasets can be quite volatile.

DPO training loss:



Run ↑	Smoothed	Value	Step	Relative
dpo-530	1.0864	0.6931	61,136	12.88 hr

Results

Below are the win rates for each of our training methods.

Method	Win-rate against baseline
SFT	40.85%
DPO	65.55%
Multi-objective DPO, 30/70 coherence-verbosity split	88.55%
Multi-objective DPO, 50/50 coherence-verbosity split	86.75%
Multi-objective DPO, 70/30 coherence-verbosity split	81.05%

Discussion

- Traditional RL methods such as SFT and DPO are solid baseline techniques to align LLM behavior with human preference.
- Multi-objective DPO offered a further boost in win-rate, highlighting the benefit of a higher quality dataset and mixing reward signals.
- A higher weight for verbosity scores over coherence scores in the dataset resulted in a more performant trained model.
- Further examination of HelpSteer dataset may be needed to draw statistically significant conclusions on the effects of different coherence-verbosity ratios.

Future Research

RL datasets are often annotated by a small, homogeneous group of experts, which can bias models toward the preferences, norms, and language styles of that group. Further, models generally optimize for broad human preference and rarely consider cultural or situational alignment, thus likely to marginalize minority perspectives. Future research could:

- Explore more drastic trade-offs such as a 10/90 split of coherence and verbosity or incorporate more signals within the reward function.
- Explore alternative scalarization methods such as
  - Pareto-optimality
  - Dynamic reweighting
- Explore further objectives beyond simple instruction-following, such as
  - Encoding ethical values
  - Encoding cultural awareness
  - Enhancing situational and contextual understanding

References

[1] https://huggingface.co/datasets/HuggingFaceTB/smoltalk

[2] https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized

[3] https://huggingface.co/datasets/Asaf-Yehudai/HelpSteer_prompt_per_row

[4] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. arXiv preprint arXiv:2305.18290, 2023.

[5] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A Roadmap to Pluralistic Alignment. arXiv:2402.05070 [cs.AI] https://arxiv.org/abs/2402.05070