DPO-*Shift*: Shifting the Distribution of Direct Preference Optimization

March 5, 2025

Xiliang Yang, Feng Jiang, Qianen Zhang, Lei Zhao, Xiao Li

The Chinese University of Hong Kong, Shenzhen

Background: Likelihood displacement issue in PO stage

Input sample: (x, y_w, y_l)

- x: Prompt
- y_w, y_l: Chosen and rejected responses, respectively

π_θ: Language Model

π_{ref}: Reference Model

DPO: $\mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(\boldsymbol{y}_w|\boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_w|\boldsymbol{x})} - \beta \log \frac{\pi_\theta(\boldsymbol{y}_l|\boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_l|\boldsymbol{x})}\right)\right]$

Likelihood displacement:

- $\log \pi_\theta(y_w|x)$ and $\log \pi_\theta(y_l|x)$ decrease simultaneously during training of DPO
- Unexpected increase in probabilities for neither preferred nor dispreferred responses
- Causes unintentional unalignment and harms the model's generalization ability

Prior works on likelihood displacement issue

Possible reasons:

- Model capacity (Tajwar et al., 2024)
- The presence of multiple training samples or output tokens (Pal et al., 2024)
- Initial SFT phase (Rafailov et al., 2024)

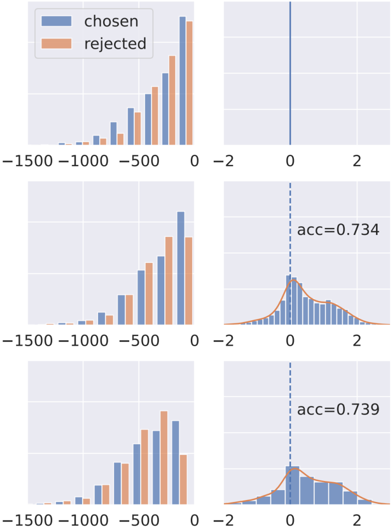
Existing solutions:

- Filter the datasets (Razin et al., 2024)

Main contributions

- Mitigate the likelihood displacement issue
  - introduce parameter function $f(\lambda)$
  - does not require modifications to dataset
- Theoretical analysis
  - Fundamental trade-off between chosen probability and reward margin between chosen and rejected responses
  - Choice strategy of $f(\lambda)$, which explicitly controls the trade-off
- Improved performance of DPO-*Shift* over DPO
  - MT-Bench scores
  - Win rate experiment

Performance: Increased chosen probability

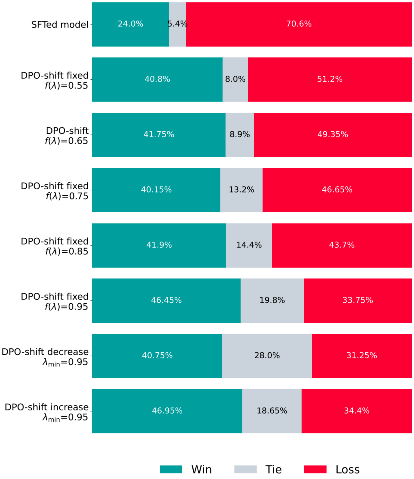


- First row: SFTed Llama 3-8B
- Second row: DPO-*Shift*
- Third row: DPO

Left: Distribution of $\log \pi_\theta(y_w|x)$ and $\log \pi_\theta(y_l|x)$
Right: KDE for the reward margin

Chosen probability is significantly improved at a minor cost of reducing reward margin compared with DPO.

Performance: Win rate experiment



- Win rate experiment against DPO using Llama 3-8B trained on the UltraFeedback dataset and tested with questions from the test split of UltraFeedback.
- When $f(\lambda)$ is closer to 1, DPO-*Shift* consistently outperforms DPO

DPO: Direct Preference Optimization

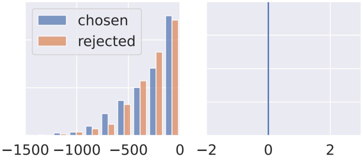


Figure: SFTed model

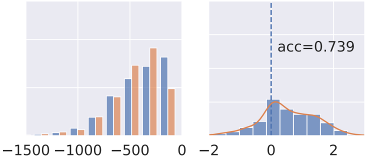
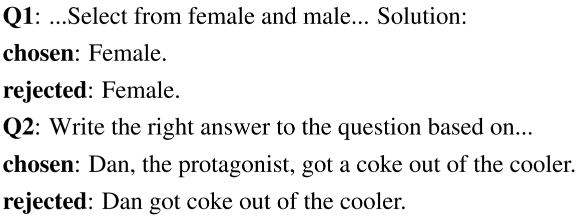


Figure: DPOed model

- Key features: No need for RL or reward model training. Matches or exceeds RLHF methods in multiple tasks.
- Two stages
  - SFT: Trains the model to maximize response likelihood using cross-entropy loss. Both responses concentrate around the high likelihood area.
  - PO: Optimizes the model to increase the probability margin between chosen and rejected responses using a reward-based objective.
- Likelihood displacement: Both probabilities of chosen and rejected responses decreased after DPO

Motivation: Cause of likelihood displacement



Q1: ...Select from female and male... Solution:

chosen: Female.

rejected: Female.

Q2: Write the right answer to the question based on...

chosen: Dan, the protagonist, got a coke out of the cooler.

rejected: Dan got coke out of the cooler.

Figure: Response examples from UltraFeedback

- Important factor: Semantic similarity between the chosen y_w and rejected y_l pairs in the dataset
- Maximize the margin $\Rightarrow$ Reduce the probability of both responses with similar semantic structures

DPO-*Shift*: Objective function

- Add a real valued function $0 < f(\lambda) < 1$ to the reward of the rejected response
- Reduce the confrontation between two semantically similar responses
- Objective function:

$$\min_{\theta} -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(\boldsymbol{y}_w|\boldsymbol{x})}{\pi_{\mathrm{ref}}(\boldsymbol{y}_w|\boldsymbol{x})} - f(\lambda)\cdot\beta\log\frac{\pi_\theta(\boldsymbol{y}_l|\boldsymbol{x})}{\pi_{\mathrm{ref}}(\boldsymbol{y}_l|\boldsymbol{x})}\right)\right]. \tag{1}$$

Evaluation indicators of PO

- Likelihood of the chosen response $\log \pi_\theta(y_w|x)$
- Indicator function of the reward margin $\mathbf{1}\{(x, y_w, y_l)| \log \frac{\pi_\theta(y_w|x)}{\pi_{\textsf{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\textsf{ref}}(y_l|x)} > 0\}$

Theoretical analysis

- Define two target functions

$$\omega_1(\theta) = \mathbb{E}\left[\log \pi_\theta\left(\boldsymbol{y}_w|\boldsymbol{x}\right)\right], \tag{2}$$

$$\omega_2(\theta) = \mathbb{E}\left[\mathbf{1}\left\{\log\frac{\pi_\theta(\boldsymbol{y}_w|\boldsymbol{x})}{\pi_{\mathsf{ref}}(\boldsymbol{y}_w|\boldsymbol{x})} - \log\frac{\pi_\theta(\boldsymbol{y}_l|\boldsymbol{x})}{\pi_{\mathsf{ref}}(\boldsymbol{y}_l|\boldsymbol{x})} > 0\right\}\right]. \tag{3}$$

Smoothed version of ω_2:

$$\omega_2(\theta) = \mathbb{E}\left[\sigma\left(\gamma\log\frac{\pi_\theta(\boldsymbol{y}_w|\boldsymbol{x})}{\pi_{\mathsf{ref}}(\boldsymbol{y}_w|\boldsymbol{x})} - \gamma\log\frac{\pi_\theta(\boldsymbol{y}_l|\boldsymbol{x})}{\pi_{\mathsf{ref}}(\boldsymbol{y}_l|\boldsymbol{x})}\right)\right], \tag{4}$$

- Gap functions after one step of update:

$$\begin{cases} g_1(t+1) = \omega_1(\theta_{t+1})\Big|_{\mathsf{DPO\text{-}Shift}} - \omega_1(\theta_{t+1})\Big|_{\mathsf{DPO}}, \\ g_2(t+1) = \omega_2(\theta_{t+1})\Big|_{\mathsf{DPO\text{-}Shift}} - \omega_2(\theta_{t+1})\Big|_{\mathsf{DPO}}. \end{cases} \tag{5}$$

Theoretical analysis

Theorem 1 Charaterization of Gap Function

Given θ_t and learning rate η and denote

$$c(\theta) = \gamma\sigma\left(f(\lambda)\cdot\gamma\log\frac{\pi_\theta(\boldsymbol{y}_l|\boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_l|\boldsymbol{x})} - \gamma\log\frac{\pi_\theta(\boldsymbol{y}_w|\boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_w|\boldsymbol{x})}\right),$$

$$\eta_1(\theta) = \eta\sigma\left(\log\frac{\pi\left(\boldsymbol{y}_l|\boldsymbol{x}\right)}{\pi_{\text{ref}}(\boldsymbol{y}_l|\boldsymbol{x})} - \log\frac{\pi\left(\boldsymbol{y}_w|\boldsymbol{x}\right)}{\pi_{\text{ref}}(\boldsymbol{y_w}|\boldsymbol{x})}\right).$$

We have

$$\begin{bmatrix} g_1(t+1) = (1-f(\lambda))u_1, \\ g_2(t+1) = (1-f(\lambda))u_2. \end{bmatrix} \tag{6}$$

Here,

$$\begin{aligned} u_1 &= \mathbb{E}\left[c(\theta)\cdot\nabla_\theta\log\pi_\theta\left(\boldsymbol{y}_l|\boldsymbol{x}\right)^\top\nabla_\theta\log\pi_\theta\left(\boldsymbol{y}_w|\boldsymbol{x}\right)\right], \\ u_2 &= \mathbb{E}\left[\eta_1(\theta)\left(\nabla_\theta\log\pi_\theta\left(\boldsymbol{y}_l|\boldsymbol{x}\right)^\top\nabla_\theta\log\pi_\theta\left(\boldsymbol{y}_w|\boldsymbol{x}\right)\right.\right. \\ &\qquad \left.\left.-\left\|\nabla_\theta\log\pi_\theta\left(\boldsymbol{y}_l|\boldsymbol{x}\right)\right\|^2\right)\right]. \end{aligned} \tag{7}$$

Fundamental trade-off

- Sample-based u_1, u_2:

$$\begin{aligned} u_1^i &= c_i(\theta) \cdot \nabla_\theta \log \pi_\theta \left(\boldsymbol{y}_l^i | \boldsymbol{x}_i \right)^\top \nabla_\theta \log \pi_\theta \left(\boldsymbol{y}_w^i | \boldsymbol{x}_i \right) \\ u_2^i &= \eta_1 \Big(\nabla_\theta \log \pi_\theta \left(\boldsymbol{y}_l^i | \boldsymbol{x}_i \right)^\top \nabla_\theta \log \pi_\theta \left(\boldsymbol{y}_w^i | \boldsymbol{x}_i \right) \\ &\qquad - \left\| \nabla_\theta \log \pi_\theta \left(\boldsymbol{y}_l^i | \boldsymbol{x}_i \right) \right\|^2 \Big). \end{aligned}$$

- Sample average of u_1, u_2:
  $u_1 = \sum_i u_1^i / |\mathcal{D}_{\text{ref}}|$, $u_2 = \sum_i u_2^i / |\mathcal{D}_{\text{ref}}|$
- 71.4% of u_1^i are positive and 81.7% of u_2^i are negative
  $\Rightarrow$ positivity of u_1 and negativity of u_2
- Trade-off:
  - Chosen probability improved:

$$0 < f(\lambda) < 1 \implies 1 - f(\lambda) > 0 \implies g_1 > 0 \text{ as } u_1 > 0.$$

  - Margin may decrease: $u_2 < 0 \Rightarrow g_2 < 0$

Indications for choosing $f(\lambda)$

- Smaller $f(\lambda)$ leads to more increase in chosen probability while a more severe drop in the reward margin.
  $\Rightarrow$ choose a relatively large $f(\lambda)$
- Choosing strategies of $f(\lambda)$
  - `fixed`: Fixed $f(\lambda)$ along the optimization process
  - `linear_increase`: $\frac{t}{T}(\lambda_{\max} - \lambda_{\min}) + \lambda_{\min}$, T: maximal iteration steps
  - `linear_decrease`: $\frac{t}{T}(\lambda_{\min} - \lambda_{\max}) + \lambda_{\max}$
- Can $f(\lambda) > 1$?
  - Theorem 1 suggests that DPO-Shift with $f(\lambda) > 1$ can lead to improvements for both the chosen probability and reward margin
  - Using $f(\lambda)$ >1 when u_1 >0 leads to catastrophic results

Verification experiment

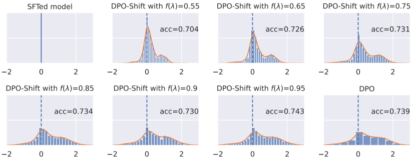


Figure: Small $f(\lambda)$:decreased reward accuracy and shifted reward margin

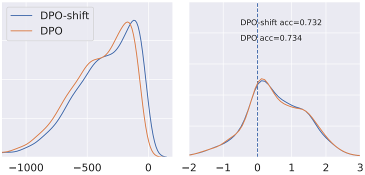


Figure: DPO-Shift with $f(\lambda) = 0.99$ and DPO

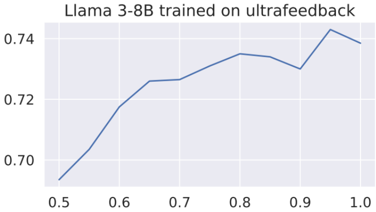


Figure: Larger $f(\lambda)$:reward accuracy increases

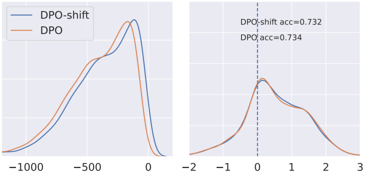
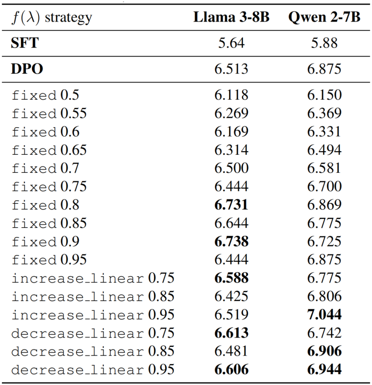


Figure: `fixed` and `linear_decrease`

MT Bench score



$f(\lambda)$ strategy	**Llama 3-8B**	**Qwen 2-7B**
SFT	5.64	5.88
DPO	6.513	6.875
`fixed` 0.5	6.118	6.150
`fixed` 0.55	6.269	6.369
`fixed` 0.6	6.169	6.331
`fixed` 0.65	6.314	6.494
`fixed` 0.7	6.500	6.581
`fixed` 0.75	6.444	6.700
`fixed` 0.8	**6.731**	6.869
`fixed` 0.85	6.644	6.775
`fixed` 0.9	**6.738**	6.725
`fixed` 0.95	6.444	6.875
`increase_linear` 0.75	**6.588**	6.775
`increase_linear` 0.85	6.425	6.806
`increase_linear` 0.95	6.519	**7.044**
`decrease_linear` 0.75	**6.613**	6.742
`decrease_linear` 0.85	6.481	**6.906**
`decrease_linear` 0.95	**6.606**	**6.944**

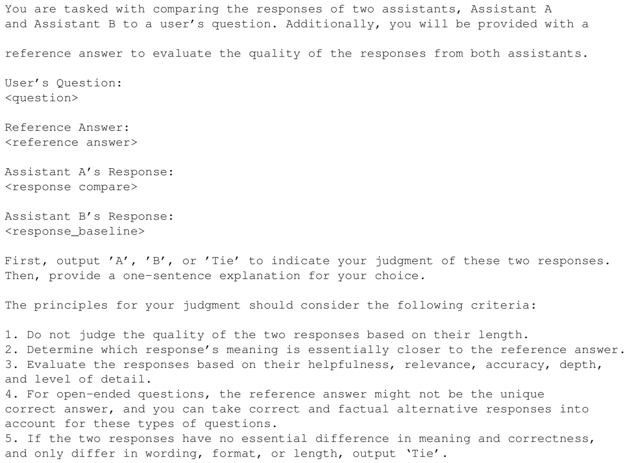
Figure: MT Bench score

- Perplexity measures the model's uncertainty about the data.
- `fixed` $f(\lambda) = 0.95$, perplexity results:
  - DPO-*Shift*: 4.475
  - DPO:18.996

Win rate experiment

Prompt design

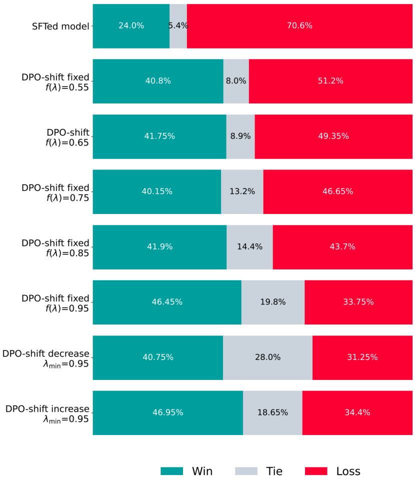
- The judge model is provided with the question, the reference answer from the dataset, and the answers generated by DPO-*Shift* and DPO.



```
You are tasked with comparing the responses of two assistants, Assistant A
and Assistant B to a user’s question. Additionally, you will be provided with a

reference answer to evaluate the quality of the responses from both assistants.

User’s Question:
<question>

Reference Answer:
<reference answer>

Assistant A’s Response:
<response compare>

Assistant B’s Response:
<response_baseline>

First, output ’A’, ’B’, or ’Tie’ to indicate your judgment of these two responses.
Then, provide a one-sentence explanation for your choice.

The principles for your judgment should consider the following criteria:

1. Do not judge the quality of the two responses based on their length.
2. Determine which response’s meaning is essentially closer to the reference answer.
3. Evaluate the responses based on their helpfulness, relevance, accuracy, depth,
and level of detail.
4. For open-ended questions, the reference answer might not be the unique
correct answer, and you can take correct and factual alternative responses into
account for these types of questions.
5. If the two responses have no essential difference in meaning and correctness,
and only differ in wording, format, or length, output ‘Tie’.
```

Figure: Llama 3.3-70B prompts for win rate experiment

Win rate experiment



- Win rate experiment against DPO using Llama 3-8B trained on the UltraFeedback dataset and tested with questions from the test split of UltraFeedback.
- When $f(\lambda)$ is closer to 1, DPO-*Shift* consistently outperforms DPO

Conclusion

Key features:

- **Likelihood displacement mitigation.** Introduces a parameter function $f(\lambda)$ in the Bradley-Terry model to adjust the rejected response's reward.
- **Identify fundamental trade-off.** Increases the probability of chosen responses leads to a decrease in the reward margin between chosen and rejected responses.
- **Consistency with win rate experiments** Validates the theoretical trade-off and the improvement of the chosen probability
- **Simplicity** Retains the simplicity of DPO compared to RLHF. Does not require additional dataset modifications or extensive computational resources.

References

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

[2] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In NeurIPS, 2023.

[3] Razin, N., Malladi, S., Bhaskar, A., Chen, D., Arora, S., and Hanin, B. Unintentional unalignment: Likelihood displacement in direct preference optimization. arXiv preprint arXiv:2410.08847, 2024.

[4] Tajwar, F., Singh, A., Sharma, A., Rafailov, R., Schneider, J., Xie, T., Ermon, S., Finn, C., and Kumar, A. Preference fine-tuning of llms should leverage suboptimal, on-policy data. arXiv preprint arXiv:2404.14367, 2024.