NIPS 2025

VPO: Reasoning Preferences Optimization Based on V-Usable Information

Zecheng Wang, Chunshan Li , Yupeng Zhang, Han Liu , Bingning Wang, Dianhui Chu , and Dianbo Sui

CONTENTS

Background

PART 01

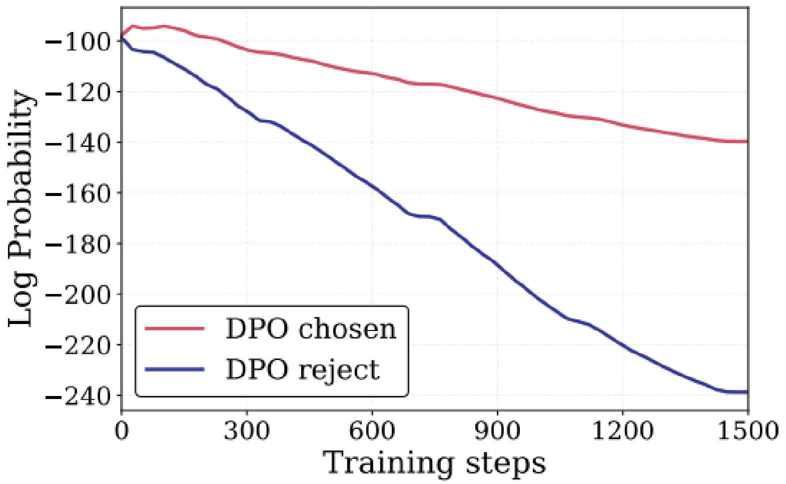
Background

Direct preference optimization (DPO):

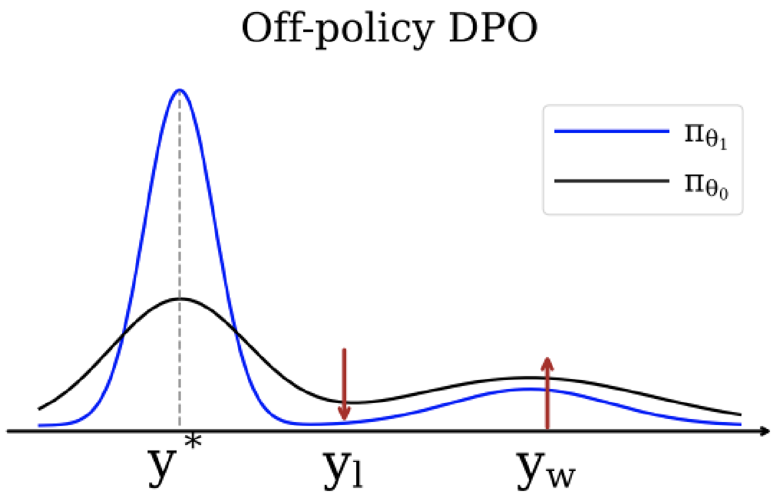
$$r(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} + \beta \log Z(x)$$

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

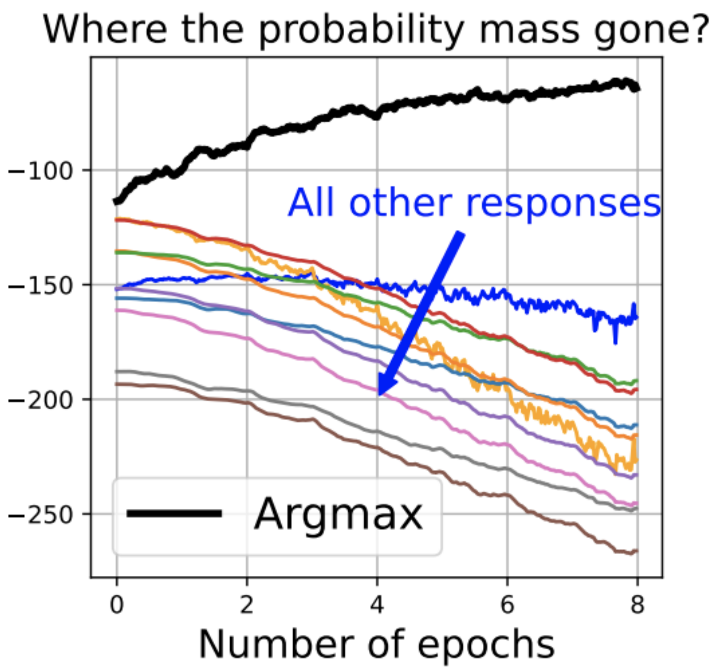
Background



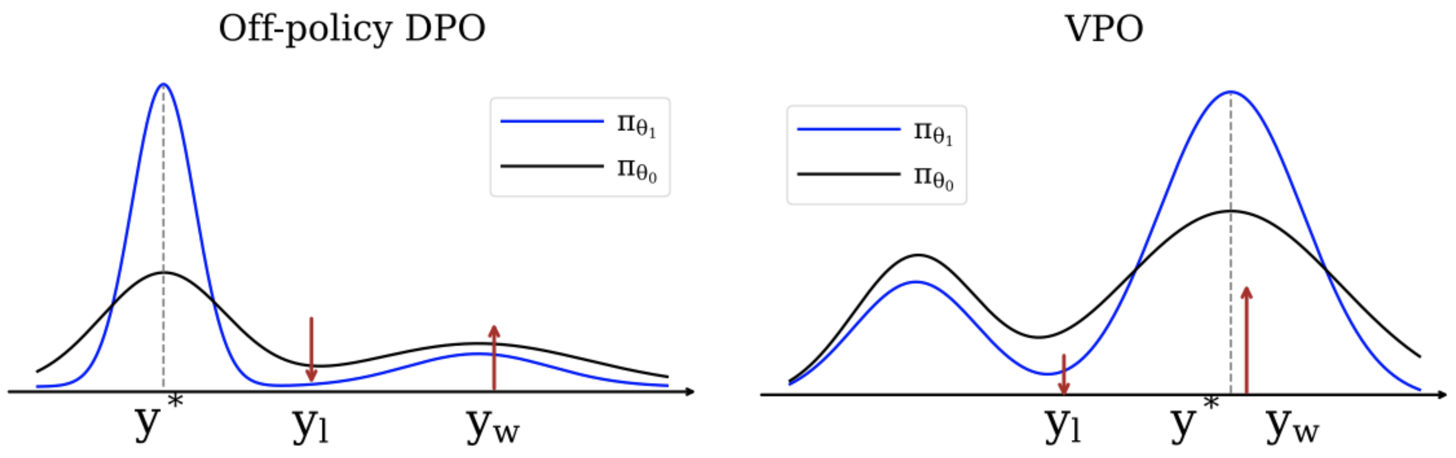
(a) Log-Likelihood decline of preference samples in DPO



(b) The squeezing effect in DPO



Background



(b) The squeezing effect in DPO

(c) Optimization performance of VPO

DPO's Limitations:

- Fixed preference data causes a distribution shift between the policy and initial model, resulting in non-uniform outputs.
- DPO's Reward does not directly align with the objective of generation (the reference model is not involved)
- DPO minimizes non-preference responses, causing non-preference samples to fall into the model's low-confidence region.

Method

PART 02

Method

Negative Gradient Constraint of DPO:

$$\mathcal{L}_{DPO_{mod}}(\pi_\theta;\pi_{ref}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - (1-v)\beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right] \tag{3}$$

$$L = -\log\sigma(r), \quad r = \beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - (1-v)\beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$$

$$\frac{\partial L}{\partial\pi_\theta(y_w|x)} = \frac{\partial L}{\partial r}\cdot\frac{\partial r}{\partial\pi_\theta(y_w|x)} = (\sigma(r)-1)\cdot\frac{\beta}{\pi_\theta(y_w|x)}$$

$$\frac{\partial L}{\partial\pi_\theta(y_l|x)} = \frac{\partial L}{\partial r}\cdot\frac{\partial r}{\partial\pi_\theta(y_l|x)} = (1-\sigma(r))\cdot\frac{\beta(1-v)}{\pi_\theta(y_l|x)}$$

Method

Negative Gradient Constraint of DPO:

Limitations:

(1) potential performance sub-optimality may be induced by static constraints.

(2) failure to adapt to sample-specific characteristics such as noise or informativeness.

Improve:

Preference and non-preference samples will mutually influence each other during DPO training.

Focusing on reasoning tasks, We characterize the correlation between texts at two levels: the token-level and the information-level.

Token-level issue: Prefix similarity; Solution path diversity

Method

VPO: Selective Negative Gradient Constraint Based on V-usable information

Conditional V-Entropy:

$$H_{\mathcal{V}}(Y|X) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log f[X](Y)]$$

V-usable information:

$$I_{\mathcal{V}}(X \to Y) = H_{\mathcal{V}}(Y \mid \varnothing) - H_{\mathcal{V}}(Y \mid X)$$

Pointwise V-usable information:

$$\mathbf{PVI}(x \to y) = -\log g[\varnothing](y) + \log g[x](y)$$

Method

VPO: Selective Negative Gradient Constraint Based on V-usable information

$$\mathbf{PVI}_l = \mathbf{PVI}(c_l \rightarrow y|x) = -\log \pi_0\,(y|x) + \log \pi_0\,(y|x, c_l)$$

$$\mathcal{L}_{VPO}\,(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x,c_w)}{\pi_{ref}(y_w|x,c_w)} - \beta(1-v)\log\frac{\pi_\theta(y_l|x,c_l)}{\pi_{ref}(y_l|x,c_l)}\right)\right]$$

$$v = \begin{cases} 0, & \mathbf{PVI}_l > 0 \\ \sigma(-\mathbf{PVI}_l), & \mathbf{PVI}_l < 0 \end{cases}$$

Experiment

PART 03

Experiment

Setup:

$$D_i^w = \{c_i^n, y_i^n, x_i^n \mid r_i^n = 1\} \quad D_i^l = \{c_i^n, y_i^n, x_i^n \mid r_i^n = 0\}$$

$$D^{pairs} = \left\{(c_i^{w_k}, y_i^{w_k}), (c_i^{l_k}, y_i^{l_k}) \,\middle|\, \forall x_i \in D \text{and } k \in [K]\right\}$$

Use Llama 3.1-8B-Base, Llama-3.1-8B Instruct, Qwen-2.5-7B-Base Qwen-2.5-7B-Instruct,

the training data constructed for each model contains 30k-40k sample pairs

Experiment

Table 1: Results of VPO, DPO and its variants on diverse mathematical reasoning tasks. The best results are highlighted in **bold**, while the second-best ones are <u>underlined</u>.

Method	Qwen2.5-7B-Base MATH 500	GSM8k	Minerva MATH	Olympiad MATH	AMC 23	Avg	Qwen2.5-7B-Instruct MATH 500	GSM8k	Minerva MATH	Olympiad MATH	AMC 23	Avg
Base	59.00	79.98	15.07	21.93	18.07	38.81	**73.20**	84.23	27.94	36.44	<u>44.58</u>	<u>53.28</u>
DPO	61.00	80.89	21.32	27.11	32.53	44.57	45.60	75.66	**28.31**	33.63	<u>44.58</u>	45.56
TDPO	59.20	79.68	17.28	26.22	28.92	42.26	48.00	77.33	23.53	20.15	34.94	40.79
SimPO	64.60	74.15	20.59	26.07	<u>33.73</u>	43.83	43.80	72.86	19.85	14.52	18.07	33.82
IPO	51.80	75.51	15.44	23.41	32.53	39.74	71.20	<u>84.99</u>	26.84	**37.19**	<u>44.58</u>	52.96
RPO	<u>66.40</u>	<u>84.46</u>	<u>21.69</u>	<u>27.26</u>	31.33	<u>46.23</u>	56.20	81.27	27.81	33.93	39.76	47.79
VPO	**68.80**	**84.91**	**23.89**	**30.52**	**45.78**	**50.78**	<u>71.60</u>	**86.73**	**28.31**	<u>36.44</u>	**48.19**	**54.26**

Method	Llama-3.1-8B-Base MATH 500	GSM8k	Minerva MATH	Olympiad MATH	AMC 23	Avg	Llama-3.1-8B-Instruct MATH 500	GSM8k	Minerva MATH	Olympiad MATH	AMC 23	Avg
Base	17.40	55.80	0.37	0.15	0.00	14.74	45.00	80.52	**22.43**	15.26	<u>27.71</u>	<u>38.18</u>
DPO	10.00	54.51	4.04	1.93	2.41	14.58	18.40	54.51	9.93	5.48	7.23	19.11
TDPO	14.80	59.29	1.47	1.33	0.00	15.38	22.75	73.09	12.50	6.52	6.02	24.18
SimPO	19.20	55.88	**8.46**	1.63	4.82	18.00	31.80	74.60	10.66	7.70	15.66	28.09
IPO	3.80	61.94	0.00	0.15	1.20	13.42	**47.20**	81.35	20.22	**15.41**	25.30	37.90
RPO	<u>19.60</u>	**65.14**	<u>7.35</u>	<u>2.37</u>	**8.43**	<u>20.58</u>	31.20	<u>81.80</u>	14.71	9.48	7.23	28.88
VPO	**20.80**	<u>63.84</u>	6.62	**3.56**	**8.43**	**20.65**	<u>46.40</u>	**83.62**	<u>20.96</u>	**15.41**	**30.12**	**39.30**

Experiment

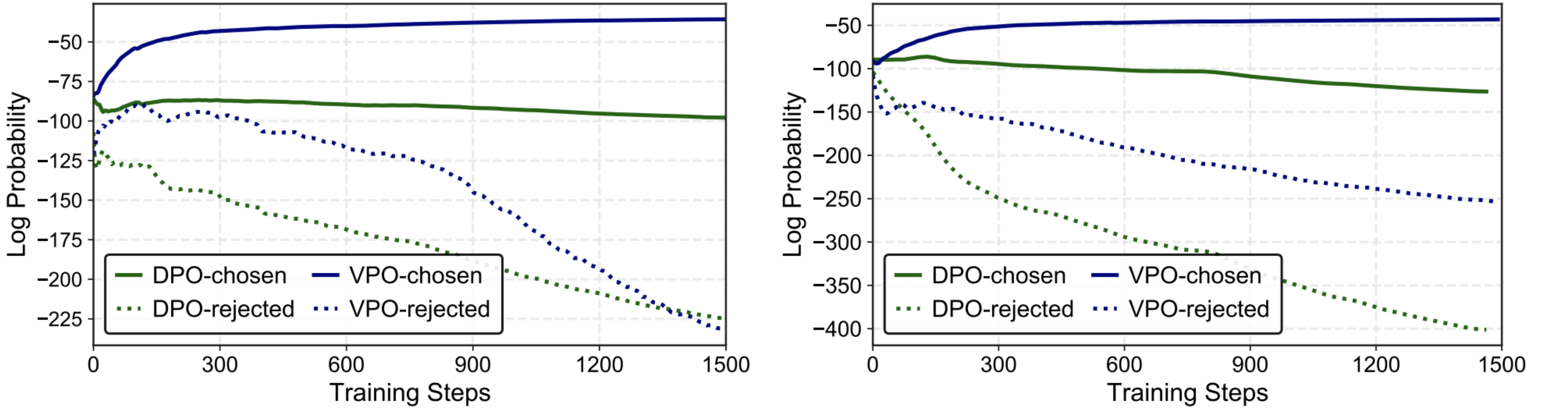


Figure 2: The log probability change curves of preference (chosen) and non-preference (rejected) samples for VPO and DPO across different models. Left: Llama-3.1-8B-Base, Right: Llama-3.1-8B-Instruct.

Experiment

Table 2: Performance comparison of DPO vs VPO across diverse math benchmarks under varying v-constraints. The best results are highlighted in **bold**, while the second-best ones are underlined.

Method	Llama-3.1-8B-Instruct						Qwen2.5-7B-Base					
	MATH 500	GSM8k	Minerva MATH	Olympiad MATH	AMC 23	Avg	MATH 500	GSM8k	Minerva MATH	Olympiad MATH	AMC 23	Avg
Base	45.00	80.52	**22.43**	15.26	27.71	38.18	59.00	79.98	15.07	21.93	18.07	38.81
DPO	18.40	54.51	9.93	5.48	7.23	19.11	61.00	80.89	21.32	27.11	32.53	47.58
0.1	22.80	64.06	13.97	6.37	6.02	22.64	67.60	84.46	22.79	29.19	31.33	**51.01**
0.2	21.80	67.55	11.76	6.81	12.05	24.00	68.60	84.84	20.59	29.04	39.76	50.77
0.3	25.00	70.74	15.44	6.96	9.64	25.56	**68.80**	84.15	20.96	29.48	38.55	50.85
0.4	31.60	76.50	15.81	9.33	10.84	28.82	68.60	83.40	20.59	29.33	43.37	50.48
0.5	34.20	74.75	16.91	11.85	12.05	29.95	68.40	83.62	21.32	28.59	39.76	50.48
0.6	39.20	76.50	17.65	12.59	18.01	32.79	66.60	85.67	20.22	27.20	40.96	49.92
0.7	44.80	79.53	18.01	13.78	14.46	34.12	65.60	**86.28**	20.59	28.89	37.35	50.34
0.8	44.40	77.18	19.12	14.07	25.30	36.01	65.80	85.97	19.49	27.56	42.17	49.70
0.9	16.40	48.90	0.74	4.41	1.20	14.33	66.80	85.37	20.22	27.56	39.76	49.99
VPO	**46.40**	**83.62**	20.96	**15.41**	**30.12**	**39.30**	**68.80**	84.91	**23.89**	**30.52**	**45.78**	**52.03**

Experiment

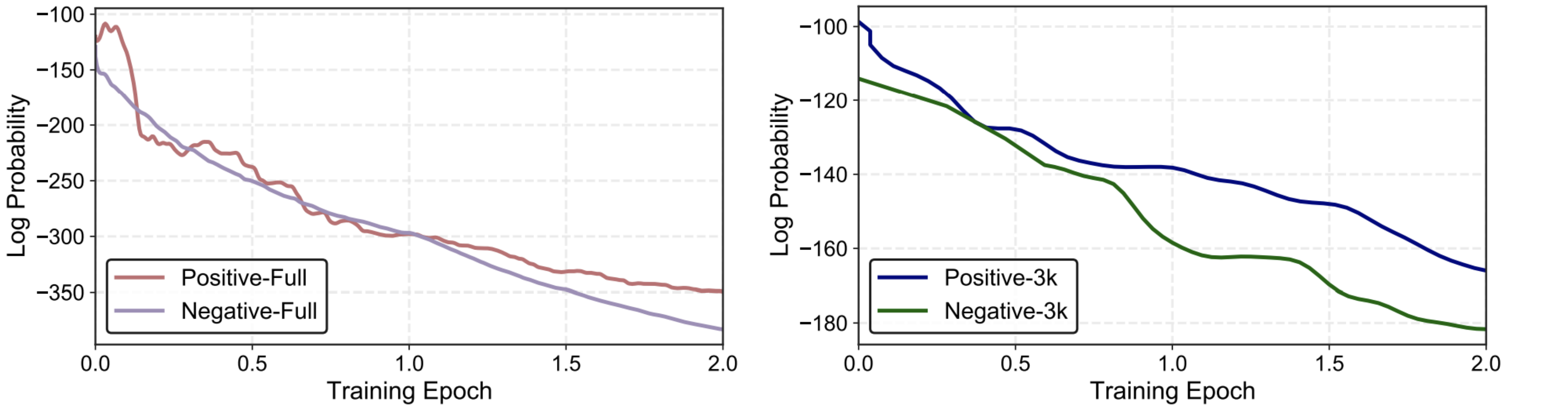


Figure 3: Decline curves of log-probabilities for non-preference samples under different configurations. Left: positive and negative non-preference samples in full training. Right: independent training on 3k preference pairs consisting exclusively of positive and negative non-preference samples.

Experiment

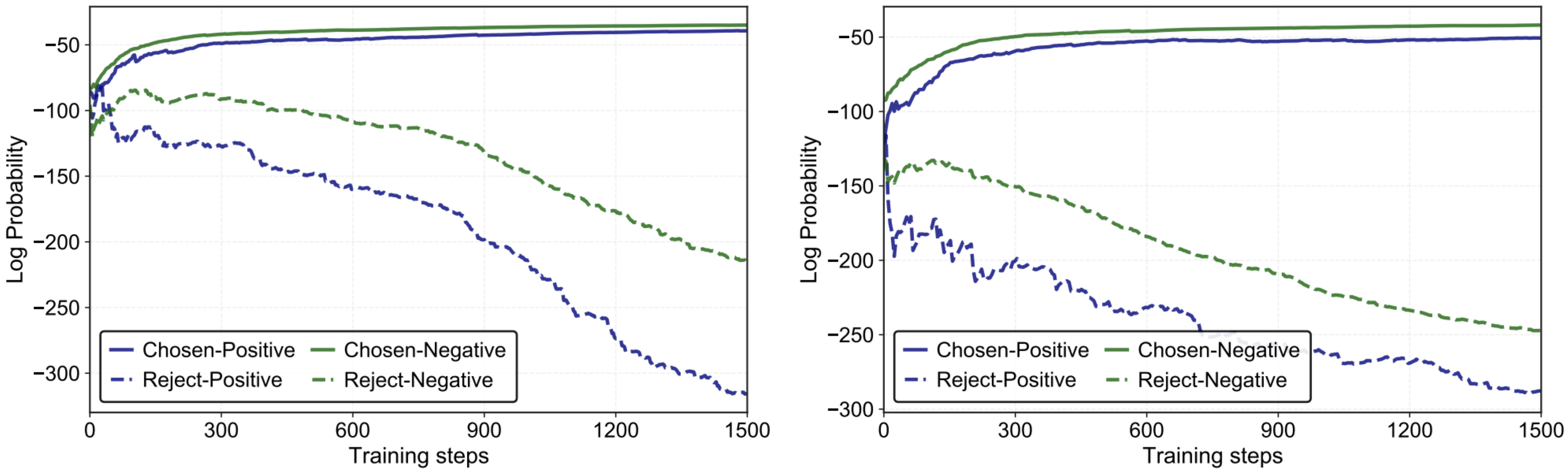


Figure 4: The log probability change curves of preference pairs under negative gradient constraint (negative) and unconstrained (positive) conditions during VPO training. Left: Llama-3.1-8B-Base, Right: Llama-3.1-8B-Instruct.

Experiment

Table 3: Results of VPO, DPO, and their variants on Qwen3-14B-Base across various mathematical reasoning tasks. The dashed line represents the Negative Gradient Constraint method based on different similarity metrics: Embedding Cosine similarity and Jaccard-based textual similarity.

Method	MATH500	GSM8k	Minerva MATH	Olympiad MATH	AMC23	AIME24	Avg
Base	63.60	93.93	24.63	21.78	22.89	0.00	37.81
DPO	76.40	94.09	28.68	33.63	45.78	20.00	49.76
TDPO	70.40	94.31	26.47	27.56	38.55	13.33	45.10
Simpo	75.60	95.75	31.25	32.44	43.37	16.67	49.18
IPO	64.60	94.24	25.00	22.81	22.89	10.00	39.92
RPO	78.00	95.68	32.35	34.67	51.81	13.33	50.97
Negative Gradient Constraint							
Cosine	75.60	95.75	29.78	34.67	44.58	20.00	50.06
Jaccard	78.00	95.98	32.35	**37.78**	49.40	16.67	51.70
VPO	**79.00**	**96.06**	**35.66**	35.41	**53.01**	**26.67**	**54.30**

Experiment

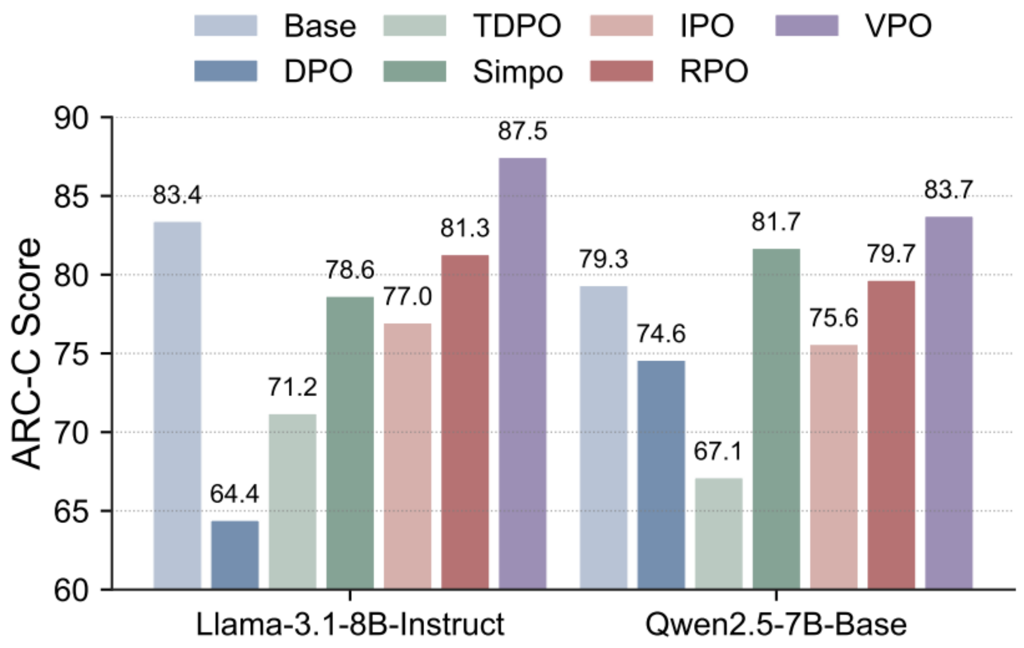


Figure 5: Performance comparison of different methods on ARC-Challenge.