ICLR

Group-Normalized Implicit Value Optimization for Language Models

Yunseon Choi, Junyoung Jang, Chaeyoung Oh, Minchan Jeong,
Doohwan Hwang, Kee-Eung Kim

KAIST

KAIST

National AI Research Lab

KL-Regularized Policy Optimization

- KL-regularized policy optimization is a widely used approach for refining language models.
- The goal is to train a policy π_θ that maximizes the expected reward while staying close to a reference policy $\pi_{\theta_{\text{old}}}$,

$$\max_\theta \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[R(x,y)-\alpha\log\frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)}\right]$$

where x is query and y is the completion generated by the policy.

- The optimal policy has a known closed-form solution:

$$\pi_{\theta^\star}(y|x)=\frac{\pi_{\theta_{\text{old}}}(y|x)\mathrm{e}^{R(x,y)/\alpha}}{Z(x)}$$

where $Z(x)=\mathbb{E}_{\pi_{\theta_{\text{old}}}(y|x)}\left[\mathrm{e}^{R(x,y)/\alpha}\right]$ is the partition function.

KL-Regularized RL with Sparse Reward

- In language tasks, the reward is sparse, typically given only upon completion of the entire generation.

Question: *Let* $T = 11$. *Compute the value of* x *that satisfies* $\sqrt{20 + \sqrt{T + x}} = 5$. x

step1: Restate the problem.

We need to find the value of x that satisfies the equation $\sqrt{20 + \sqrt{T + x}} = 5$, where $T = 11$.

step2: Substitute $T = 11$ into the equation.

The equation becomes $\sqrt{20 + \sqrt{11 + x}} = 5$.

$y_{<t}$

$$\pi_{\theta^\star}(y_{<t}|x) \propto \pi_{\theta_{\text{old}}}(y_{<t}|x)\ \{??\}$$

step7: Verify the solution by substituting $x = 14$ back into the original equation.

$$\sqrt{20 + \sqrt{11 + 14}} = \sqrt{20 + \sqrt{25}} = \sqrt{20 + 5} = \sqrt{25} = 5$$

The solution satisfies the original equation.

Therefore, the value of x is $\boxed{14}$.

y

$$\pi_{\theta^\star}(y|x) \propto \pi_{\theta_{\text{old}}}(y|x)\mathrm{e}^{R(x,y)/\alpha}$$

Toward Step-level Value

- *(Theorem 1)* Suppose that $\pi_{\theta^\star}$ and $\pi_{\theta_{\text{old}}}$ are auto-regressive policies, and that they satisfy the optimal condition for complete response y. For any t, $\pi_{\theta^\star}$ satisfies:

$$\pi_{\theta^\star}(y_{<t}|x) = \frac{\pi_{\theta_{\text{old}}}(y_{<t}|x)\mathrm{e}^{V(x,y_{<t})}}{Z(x)}$$

where a soft value function $V(x, y_{<t})$, which represents the expected future reward from $y_{<t}$, is defined as:

$$V(x, y_{<t}) := \begin{cases} R(x,y)/\alpha & t = T, \\ \log \mathbb{E}_{\pi_{\theta_{\text{old}}}(y|y_{<t},x)}\left[\mathrm{e}^{R(x,y)/\alpha}\right] & t < T. \end{cases}$$

Toward Critic-free Algorithm

- While the soft value function can be modeled with an explicit network, we pursue a direct policy objective that avoids explicit value modeling.
- We can express the value function implicitly in terms of a policy ratio:

$$V(x, y_{<t}) = \log Z(x) + \log \frac{\pi_{\theta^\star}(y_{<t}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}|x)}$$

- The mean-squared error (MSE) between the policy-defined value and the true value function can be an objective for direct policy training loss.

$$\mathcal{L}_{\text{MSE}}(Z, \theta) = \left(\boxed{\log Z(x)} + \log \frac{\pi_\theta(y_{<t}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}|x)} - \log \mathbb{E}_{y \sim \pi_{\theta_{\text{old}}}(\cdot|y_{<t}, x)} \left[e^{R(x,y)/\alpha} \right] \right)^2$$

→ A key challenge remains: the partition function still requires an auxiliary network to approximate.

The Group-normalized objective for V_θ

- Sample a group of K completions $\{y^{(i)}\}_{i=0}^{K-1}$ from $\pi_{old}(\cdot\,|x)$,

$$\min_{V_\theta} \mathbb{E}_{\substack{x\sim\mathcal{D},t\sim U\{1,\ldots,T\},\\ y^{(0:K-1)},\, y_{<t}^{(0:K-1)}\sim\pi_{\text{old}}(\cdot\,|\,x)}} \left[-\sum_{i=0}^{K-1} e^{R\left(x,y^{(i)}\right)/\alpha} \log \frac{e^{V_\theta\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=0}^{K-1} e^{V_\theta\left(x,y_{<t}^{(j)}\right)}} \right]$$

The Group-normalized objective for V_θ

- Sample a group of K completions $\{y^{(i)}\}_{i=0}^{K-1}$ from $\pi_{old}(\cdot\,|x)$,

$$\min_{V_\theta} \mathbb{E}_{\substack{x\sim\mathcal{D},t\sim U\{1,\ldots,T\},\\ y^{(0:K-1)},\, y_{<t}^{(0:K-1)}\sim\pi_{\text{old}}(\cdot\,|\,x)}} \left[-\sum_{i=0}^{K-1} e^{R\left(x,y^{(i)}\right)/\alpha} \log \frac{e^{V_\theta\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=0}^{K-1} e^{V_\theta\left(x,y_{<t}^{(j)}\right)}} \right]$$

joint

$$\pi_{\text{old}}\left(y_{<t}^{(0:K-1)}\,\middle|\,x\right)\pi_{\text{old}}\left(y^{(0:K-1)}\,\middle|\,x,y_{<t}^{(0:K-1)}\right)$$

$$\mathbb{E}_{\pi_{\text{old}}\left(y^{(i)}\,\middle|\,x,y_{<t}^{(i)}\right)}\left[e^{R\left(x,y^{(i)}\right)/\alpha}\right]\left(=e^{V\left(x,y^{(i)}\right)}\right)$$

The Group-normalized objective for V_ψ

- Sample a group of K completions $\{y^{(i)}\}_{i=0}^{K-1}$ from $\pi_{old}(\cdot\,|x)$,

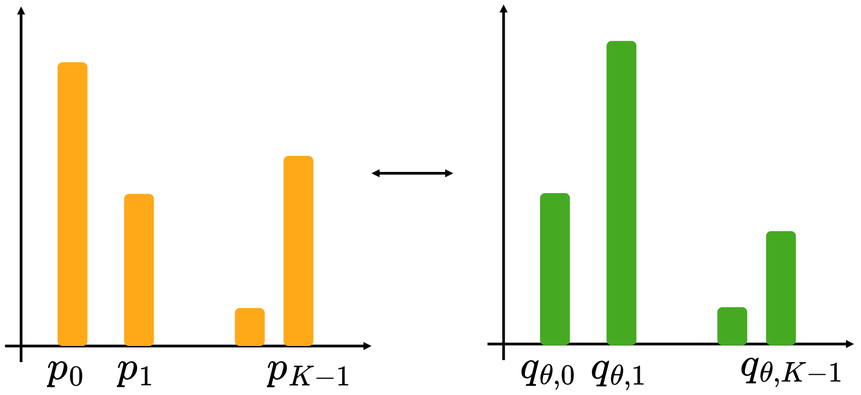
$$\min_{V_\theta} \mathbb{E}_{\substack{x\sim\mathcal{D},t\sim U\{1,\ldots,T\},\\ y_{<t}^{(0:K-1)}\sim\pi_{\text{old}}(\cdot\,|\,x)}} \left[\left(\sum_{j=1}^{K-1} e^{V\left(x,y_{<t}^{(j)}\right)} \right) \cdot - \left(\sum_{i=0}^{K-1} \frac{e^{V\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=1}^{K-1} e^{V\left(x,y_{<t}^{(j)}\right)}} \log \frac{e^{V_\theta\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=0}^{K-1} e^{V_\theta\left(x,y_{<t}^{(j)}\right)}} \right) \right]$$

The Group-normalized objective for V_ψ

$$\min_{V_\theta} \mathbb{E}_{\substack{x\sim\mathcal{D},t\sim U\{1,\ldots,T\},\\ y_{<t}^{(0:K-1)}\sim\pi_{\text{old}}(\cdot\mid x)}} \left[\left(\sum_{j=1}^{K-1} e^{V\left(x,y_{<t}^{(j)}\right)} \right) \cdot - \left(\sum_{i=0}^{K-1} \underbrace{\frac{e^{V\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=1}^{K-1} e^{V\left(x,y_{<t}^{(j)}\right)}}}_{p_i} \log \underbrace{\frac{e^{V_\theta\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=0}^{K-1} e^{V_\theta\left(x,y_{<t}^{(j)}\right)}}}_{q_{\theta,i}} \right) \right]$$

Cross-Entropy Loss

$$(p_i = q_{\theta,i} \quad \text{for}\, \forall\, i)$$



$$\frac{e^{V\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=1}^{K-1} e^{V\left(x,y_{<t}^{(j)}\right)}} = \frac{e^{V_\theta\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=0}^{K-1} e^{V_\theta\left(x,y_{<t}^{(j)}\right)}}$$

$$\rightarrow \frac{e^{V_\theta\left(x,y_{<t}^{(i)}\right)}}{e^{V\left(x,y_{<t}^{(i)}\right)}} = \frac{\sum_{j=0}^{K-1} e^{V_\theta\left(x,y_{<t}^{(j)}\right)}}{\sum_{j=1}^{K-1} e^{V\left(x,y_{<t}^{(j)}\right)}} = C_t\left(x\right)$$

The Group-normalized objective

- *(Theorem 2, Consistency up to constant shift)* Assume unlimited model capacity and data. For any $K > 1$ and $t \in \{1, \cdots, T\}$, the minimizer V_{θ^*} of *the group normalized objective* recover the soft value function V up to an additive, $y_{<t}$-independent offset $C_t(x)$:

$$V_{\theta^\star}(x, y_{<t}) = V(x, y_{<t}) + \log C_t(x),$$

equivalently, $e^{V_{\theta^*}(x, y_{<t})} = C_t(x) e^{V(x, y_{<t})}$.

- *(Corollary 3, Policy invariance to additive shifts)* For any positive scalar $C_t(x)$, let $\mathrm{V}'(\mathrm{x}, \mathrm{y}_{<t}) = V(x, y_{<t}) + \log C_t(x)$. The optimal policy for V' remains the same as the optimal policy for V.

From Theorem 1, $\left(\pi_{\theta^\star}(y_{<t} \mid x) = \frac{\pi_{\theta_{\text{old}}}(y_{<t} \mid x) e^{V(x, y_{<t})}}{Z(x)} \right)$

Group-normalized Implicit Value Optimization

- For $e^{V_{\theta^*}(x,y_{<t})} = C_t(x)e^{V(x,y_{<t})}$,

$$e^{V_{\theta^\star}(x,y_{<t})} = C_t(x)Z(x)\frac{\pi_{\theta^\star}(y_{<t}\mid x)}{\pi_{\theta_{\text{old}}}(y_{<t}\mid x)}$$

$$\min_{V_\theta} \mathbb{E}_{\substack{x\sim\mathcal{D},t\sim U\{1,\ldots,T\},\\ y^{(0:K-1)},\, y_{<t}^{(0:K-1)}\sim\pi_{\text{old}}(\cdot\mid x)}} \left[-\sum_{i=0}^{K-1} e^{R\left(x,y^{(i)}\right)/\alpha} \log \frac{e^{V_\theta\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=0}^{K-1} e^{V_\theta\left(x,y_{<t}^{(j)}\right)}} \right]$$

$$\longrightarrow \frac{\cancel{C_t(x)Z(x)}\,\frac{\pi_\theta(y_{<t}^{(i)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(i)}|x)}}{\sum_{j=0}^{K-1}\cancel{C_t(x)Z(x)}\,\frac{\pi_\theta(y_{<t}^{(j)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(j)}|x)}} = \frac{\frac{\pi_\theta(y_{<t}^{(i)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(i)}|x)}}{\sum_{j=0}^{K-1}\frac{\pi_\theta(y_{<t}^{(j)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(j)}|x)}}$$

Group-normalized Implicit Value Optimization

$$\mathcal{L}_{GN-IVO}(\theta) = \mathbb{E}_{\substack{x\sim\mathcal{D},t\sim U\{1,\ldots,T\},\\ y_{<t}^{(0:K-1)}\sim\pi_{\text{old}}(\cdot\mid x)}}\left[-\sum_{i=0}^{K-1} e^{R(x,y^{(i)})/\alpha}\left(\log\frac{\pi_\theta\left(y_{<t}^{(i)}\,\middle|\,x\right)}{\pi_{\theta_{\text{old}}}\left(y_{<t}^{(i)}\,\middle|\,x\right)} - \log\sum_{j=0}^{K-1}\frac{\pi_\theta\left(y_{<t}^{(j)}\,\middle|\,x\right)}{\pi_{\theta_{\text{old}}}\left(y_{<t}^{(j)}\,\middle|\,x\right)}\right)\right]$$

Algorithm 1 Group-Normalized Implicit Value Optimization

Input: Reward function R, learning rate η, the policy π_θ and set $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
for iterations **do**
  Sample a query $x \sim \mathcal{D}$ and generate K responses $y^{(0:K)} \sim \pi_{\theta_{\text{old}}}(\cdot|x)$
  Evaluate reward $R(x, y^{(i)})$ for all $i \in \{0, \ldots, K-1\}$
  Update θ by optimizing the following loss in Eq. 9, $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{\text{GN-IVO}}(\theta)$
  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
end for

Experiments – Baseline algorithms

- *SFT-winning* utilizes two responses for each query and performs supervised fine-tuning on the response with the higher reward.
- *Online DPO* extends DPO to an online setting.

$$\log \pi_\theta \left(y_w \mid x \right)$$

$$\log \sigma \left(\alpha \log \frac{\pi_\theta \left(y_w \mid x \right)}{\pi_{\theta_{\text{old}}} \left(y_w \mid x \right)} - \alpha \log \frac{\pi_\theta \left(y_l \mid x \right)}{\pi_{\theta_{\text{old}}} \left(y_l \mid x \right)} \right)$$

Experiments – Baseline algorithms

- *SFT-winning* utilizes two responses for each query and performs supervised fine-tuning on the response with the higher reward.
- *Online DPO* extends DPO to an online setting.
- *PPO* generates a single completion for each query and performs clipped policy-gradient updates using advantages computed by a critic network.
- *DRO* is a soft Q-learning algorithm adapted to the bandit setting.

(MSE Loss)

$$\mathcal{L}_{\mathrm{DRO}}(\psi,\theta)=\left(\log Z_\psi(x)+\log\frac{\pi_\theta(y\mid x)}{\pi_{\theta_{\mathrm{old}}}(y\mid x)}-R(x,y)/\alpha\right)^2$$

$$\nabla_\theta\mathcal{L}_{\mathrm{DRO}}(\psi,\theta)=-(R(x,y)-\log Z_\psi(x))\nabla_\theta\log\pi_\theta(y\mid x)-\frac{\alpha}{2}\nabla_\theta\left(\log\frac{\pi_\theta(y\mid x)}{\pi_{\theta_{\mathrm{old}}}(y\mid x)}\right)^2$$

(Policy Gradient)

Experiments – Baseline algorithms

- *SFT-winning* utilizes two responses for each query and performs supervised fine-tuning on the response with the higher reward.
- *Online DPO* extends DPO to an online setting.
- *PPO* generates a single completion for each query and performs clipped policy-gradient updates using advantages computed by a critic network.
- *DRO* is a soft Q-learning algorithm adapted to the bandit setting.
- *OREO* is a sequential extension of DRO that uses soft Q-learning to train a step-level value network.

(MSE Loss)
$$\mathcal{L}_{\mathrm{OREO}}(\psi,\theta)=\left(\log V_\psi\left(x,y_{<t}\right)+\log\frac{\pi_\theta\left(y\,|y_{<t},x\right)}{\pi_{\theta_{\mathrm{old}}}\left(y\,|y_{<t},x\right)}-R\left(x,y\right)/\alpha\right)^2$$

$$\nabla_\theta\mathcal{L}_{\mathrm{OREO}}(\psi,\theta)=-\left(R\left(x,y\right)-\log V_\psi\left(x,y_{<t}\right)\right)\nabla_\theta\log\pi_\theta\left(y|\,y_{<t},x\right)-\frac{\alpha}{2}\nabla_\theta\left(\log\frac{\pi_\theta\left(y\,|\,x,y_{<t}\right)}{\pi_{\theta_{\mathrm{old}}}\left(y\,|\,x,y_{<t}\right)}\right)^2$$

(Policy Gradient)

Experiments – Baseline algorithms

- *SFT-winning* utilizes two responses for each query and performs supervised fine-tuning on the response with the higher reward.
- *Online DPO* extends DPO to an online setting.
- *PPO* generates a single completion for each query and performs clipped policy-gradient updates using advantages computed by a critic network.
- *DRO* is a soft Q-learning algorithm adapted to the bandit setting.
- *OREO* is a sequential extension of DRO that uses soft Q-learning to train a step-level value network.
- *RLOO* is a REINFORCE-style policy gradient method that employs a leave-one-out estimation for advantages.

$$\frac{K}{K-1}\left(R\left(x,y\right)-\mu_{\text{group}}\right)\nabla_\theta \log \pi_\theta(y\,|\,x)$$

- *GRPO* is a PPO-style method that uses the group mean as a baseline for advantage estimation.

$$\left(\frac{R\left(x,y\right)-\mu_{\text{group}}}{\text{std}_{\text{group}}}\right)\nabla_\theta \log \pi_\theta\left(y\,|\,x\right)$$

Experiments – Mathematical Reasoning

Table 1. Comparison of our method against baselines on the math reasoning task. The Pass@3 (P@3) metric is calculated over three trials per query. Bold and underline indicate the best and second-best results, respectively.

Method	AMC2023		Minerva Math		Olympiad-Bench		AIME2024		AIME2025	
	P@1	P@3	P@1	P@3	P@1	P@3	P@1	P@3	P@1	P@3
Llama-3.1-8B-Instruct	27.5	37.5	25.7	32.7	15.6	24.2	3.3	10.0	0.0	0.0
SFT-winning	27.5	35.0	24.2	35.6	16.0	27.1	6.6	6.6	0.0	0.0
Online DPO	22.5	33.1	25.3	30.5	15.1	26.2	3.3	13.3	0.0	0.0
PPO	25.0	35.0	21.7	34.9	15.7	26.2	3.3	**16.6**	**3.3**	0.0
DRO	22.5	35.0	23.1	33.8	15.5	25.6	3.3	0.0	0.0	**6.6**
OREO	27.5	32.5	25.7	35.3	15.7	26.8	3.3	6.6	0.0	**6.6**
RLOO	35.0	40.0	**26.1**	34.1	17.9	26.1	6.6	**16.6**	0.0	0.0
GRPO	35.0	37.5	25.3	35.3	**18.8**	25.3	6.6	**16.6**	**3.3**	0.0
Ours	**42.5**	**45.0**	**26.1**	**36.0**	17.3	**27.8**	**10.0**	**16.6**	**3.3**	3.3
Qwen2.5-Math-7B	52.5	70.0	27.0	36.0	37.7	46.2	23.0	26.6	6.6	13.3
SFT-winning	57.5	62.5	30.5	36.0	38.8	49.0	23.3	26.6	**13.3**	13.3
Online DPO	57.5	70.0	27.2	37.1	36.2	49.2	23.3	30.0	10.0	13.3
PPO	47.5	67.5	28.3	37.8	38.6	49.0	23.3	30.0	10.0	13.3
DRO	55.0	67.5	31.2	37.1	37.7	48.7	23.3	33.3	10.0	13.3
OREO	55.0	70.0	**31.6**	38.6	38.5	49.1	16.6	30.0	10.0	13.3
RLOO	57.5	72.5	30.1	40.4	38.8	48.8	23.3	36.6	**13.3**	16.6
GRPO	60.0	70.0	29.7	37.8	39.2	**49.4**	26.6	33.3	6.6	13.3
Ours	**62.5**	**75.0**	**31.6**	**41.9**	**39.8**	49.0	**30.0**	**40.0**	**13.3**	**23.3**

Experiments – Text generation

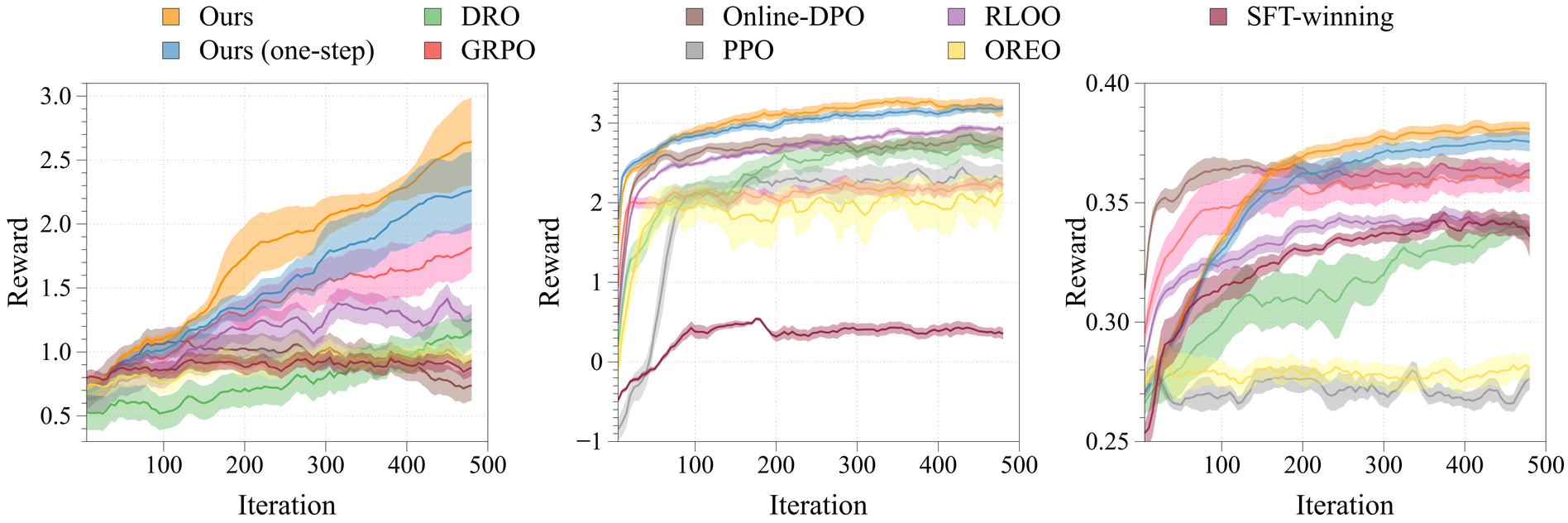


Figure 1. Training curves for our methods and baselines on the Llama-3.2-3B-Instruct model. The solid lines represent the mean reward, while the shaded regions indicate the standard deviation calculated over three random seeds.

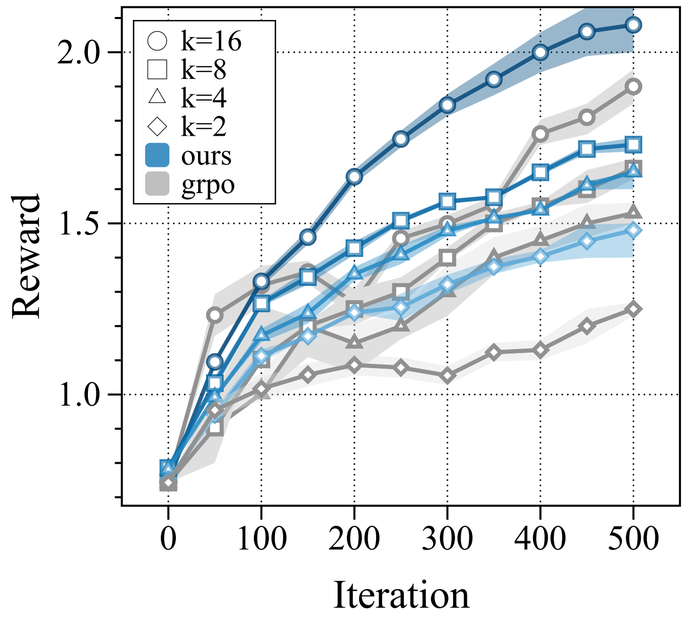
Experiments – Text generation

Table 2. Comparison of our methods against baselines on three text generation tasks. We run three random seeds and report the best scores on the test set across seeds. For Avg@3, we report the mean with std. over 3 random seeds. GM denotes the geometric mean of Avg@1 across the three tasks. Bold and underline indicate the best and the second-best accuracy for each task, respectively.

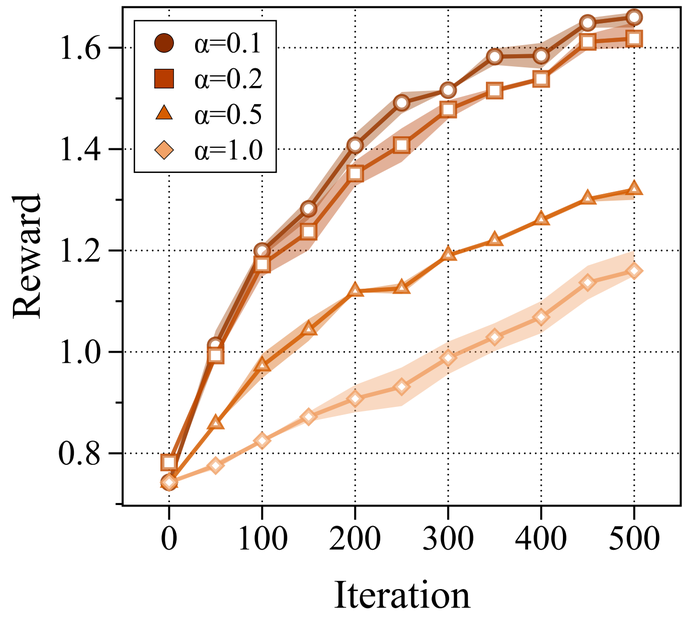
Method	Helpful assistant		TL;DR summarization		Prompt generation		GM
	Avg@1	Avg@3	Avg@1	Avg@3	Avg@1	Avg@3	
Qwen2.5-1.5B-Instruct							
SFT-winning	0.375	0.375 (0.001)	-0.710	-0.729 (0.008)	0.342	0.329 (0.004)	-0.450
Online DPO	1.271	1.260 (0.005)	0.998	0.660 (0.016)	<u>0.381</u>	<u>0.381</u> (0.000)	0.786
PPO	0.875	0.846 (0.003)	1.393	1.320 (0.015)	0.274	0.269 (0.005)	0.694
DRO	1.174	1.146 (0.007)	1.831	1.725 (0.015)	0.369	0.363 (0.004)	0.926
OREO	0.721	0.733 (0.002)	1.633	1.362 (0.033)	0.261	0.259 (0.003)	0.675
RLOO	1.120	1.043 (0.001)	**2.466**	**2.423** (0.003)	0.319	0.311 (0.005)	1.004
GRPO	<u>1.594</u>	<u>1.506</u> (0.014)	1.151	1.086 (0.021)	0.367	0.367 (0.000)	0.876
Ours (one-step)	1.446	1.389 (0.005)	2.148	2.108 (0.024)	0.379	0.372 (0.001)	<u>1.056</u>
Ours	**1.650**	**1.628** (0.009)	<u>2.418</u>	<u>2.359</u> (0.024)	**0.383**	**0.382** (0.001)	**1.152**
Llama-3.2-3B-Instruct							
SFT-winning	0.819	0.819 (0.010)	0.603	0.897 (0.011)	0.354	0.354 (0.001)	0.640
Online DPO	1.064	0.974 (0.008)	2.907	2.884 (0.005)	<u>0.372</u>	<u>0.372</u> (0.000)	1.066
PPO	0.867	0.854 (0.013)	2.807	2.748 (0.016)	0.287	0.285 (0.002)	0.887
DRO	1.317	1.295 (0.002)	3.104	3.099 (0.008)	0.350	0.345 (0.002)	1.082
OREO	0.881	0.864 (0.008)	2.715	2.649 (0.018)	0.291	0.288 (0.005)	0.886
RLOO	1.528	1.262 (0.004)	3.181	3.116 (0.008)	0.349	0.345 (0.002)	1.193
GRPO	2.013	1.996 (0.013)	2.337	2.316 (0.001)	0.358	0.346 (0.005)	1.154
Ours (one-step)	<u>2.562</u>	<u>2.555</u> (0.002)	**3.398**	**3.334** (0.007)	0.371	0.367 (0.001)	<u>1.478</u>
Ours	**3.370**	**3.281** (0.005)	<u>3.347</u>	<u>3.330</u> (0.013)	**0.384**	**0.382** (0.001)	**1.630**

Experiments – Analysis on hyperparameter

Figure 1. Analysis on hyperparameter using the Qwen2.5-1.5B-Instruct on the Anthropic HH-RLHF dataset.



(a) Group Size K



(b) Temperature Coefficient α

$$-\sum_{i=0}^{K-1} e^{R(x,y^{(i)})/\alpha} \left(\log \frac{\pi_\theta\left(y_{<t}^{(i)} \middle| x\right)}{\pi_{\theta_{\text{old}}}\left(y_{<t}^{(i)} \middle| x\right)} - \log \sum_{j=0}^{K-1} \frac{\pi_\theta\left(y_{<t}^{(j)} \middle| x\right)}{\pi_{\theta_{\text{old}}}\left(y_{<t}^{(j)} \middle| x\right)} \right)$$