# Preference Distillation via Value based Reinforcement Learning

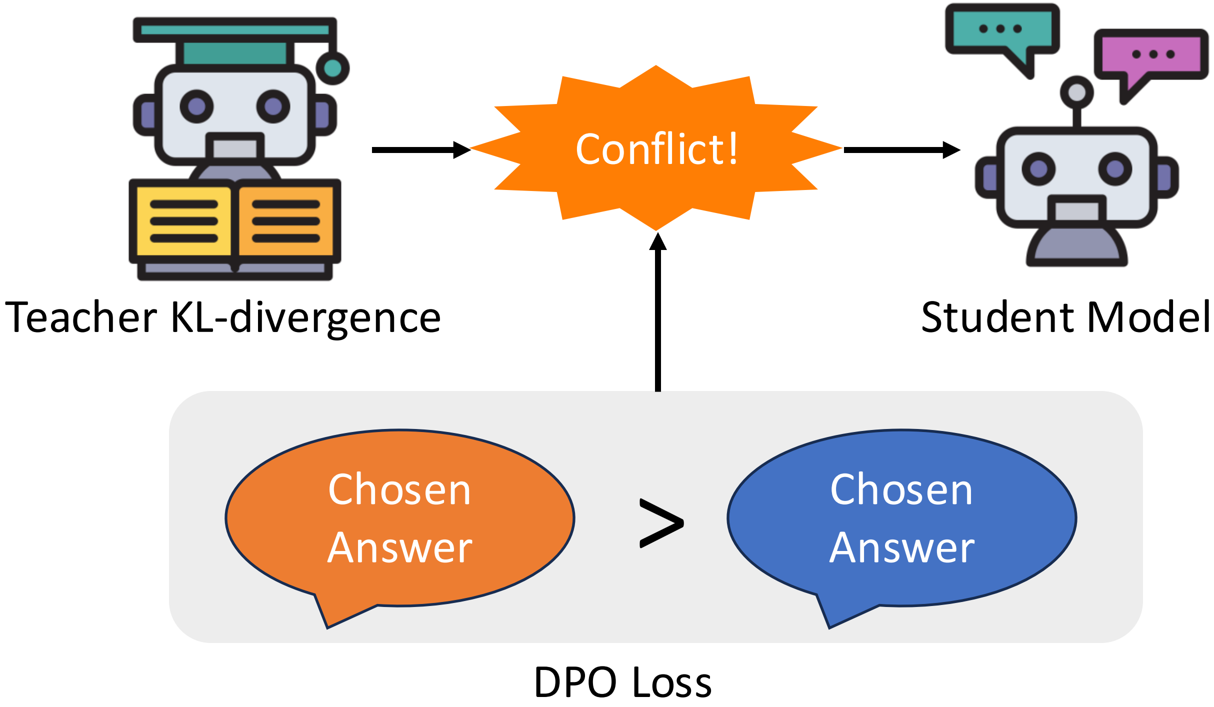
Minchan Kwon, Junwon Ko, Kangil Kim, Junmo Kim

Statistical Inference and Information Theory Laboratory

# Introduction

- **Preference Optimization, led by DPO, is becoming a key component of LLM training**
- DPO-based Knowledge Distillation is actively explored in recent research
- Compare to other RL methods, DPO enables easier dataset construction for distillation
  - You need only which is good or bad!
- However, simply adding a KL-divergence term to DPO is insufficient

# Introduction




- **In the Rense of Q-learning, KL-divergence and DPO Loss can conflict**
- **We analyze the conflict and suggest the solution that can transfer teacher's knowledge without conflict with DPO loss**

# Optimal Policy Invariance under reward shaping

- In Andrew Y. Ng’s seminal paper, it was proven that adding or subtracting a potential-based function to the reward, where the function depends only on the state, preserves the optimal policy.

$$\psi(\mathbf{s}_t, a_t) = V_\phi(\mathbf{s}_{t+1}) - V_\phi(\mathbf{s}_t).$$

- Here, V denotes a function that depends only on the state, not the action.

# Action-based Shaping Breaks Policy Invariance

- Conversely, we prove that action-dependent reward shaping breaks policy invariance.

**Corollary 2.1 (Action-based Shaping Breaks Policy Invariance).** *Let* $r'(s,a) = r(s,a) + \psi(s,a)$*, where* $\psi(s,a)$ *is composed of the function dependent with both state and action. If* $\psi(s,a) \in \{\alpha \log \pi_\phi(a \mid s), \alpha Q_\phi(s,a)\}$*, then the resulting reward violates the policy invariance guarantee and may alter the optimal policy.*

- This theoretically shows that functions used in previous distillation losses (e.g., KL-divergence) can interfere with the original reward in DPO.

# Teacher-value based knowledge distillation

- To address this, we extract the teacher's value function and incorporate it into the reward in a way that satisfies potential-based reward shaping (PBRS), ensuring policy invariance during student learning.
- We first propose a method to estimate the value function of a DPO-trained model (Lemma 1).

**Lemma 1 (Soft value function of a DPO-trained policy [25]).** *Let $Q_\phi(s, a)$ denote the token-level logits of a DPO-trained model $\pi_\phi$, and let $\beta > 0$ be the temperature of the Boltzmann policy. Then the soft value function is given by:*

$$V_\phi(s) = \beta \log \sum_{a \in \mathcal{V}} \exp(Q_\phi(s, a)/\beta).$$

# Teacher-value based knowledge distillation

- Using the teacher's value function, we augment the reward function toward a policy-invariant direction.

$$\max_{\pi_\theta} \mathbb{E}_{(s,a)\sim\pi_\theta}\left[r(s,a) + \alpha\, \psi_\phi(s,a)\right]$$

The resulting formulation can be analytically derived as follows: (show formula here)

$$\mathcal{L}_{\text{TVKD}}(\pi_\theta, \mathcal{D}; \pi_\phi) = -\mathbb{E}_{(\tau_w,\tau_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta \log \frac{\pi_\theta(a^w \mid s^w)}{\exp\left(\frac{\alpha}{\beta}\psi_\phi(s^w, a^w)\right)} - \beta \log \frac{\pi_\theta(a^l \mid s^l)}{\exp\left(\frac{\alpha}{\beta}\psi_\phi(s^l, a^l)\right)}\right)\right]$$

# Experiment

| Method | DPOMIX | | | | Helpsteer2 | | | |
|---|---|---|---|---|---|---|---|---|
| | RM(↑) | MT(↑) | AE(↑) | OLL(↑) | RM(↑) | MT(↑) | AE(↑) | OLL(↑) |
| DPO Teacher | -0.77 | 5.74 | 73.12 | 53.45 | -0.88 | 5.63 | 72.87 | 57.95 |
| SFT Teacher | -0.94 | 5.58 | 71.02 | 53.15 | -1.23 | 5.58 | 70.31 | 53.15 |
| SFT | -1.55 | 3.56 | 43.62 | 40.18 | -1.78 | 3.43 | 45.49 | 41.06 |
| DPO | -1.41 | 3.70 | 47.03 | 35.48 | -1.22 | 3.77 | 44.5 | 41.24 |
| SimPO | -1.22 | 3.65 | 46.12 | 41.5 | -1.37 | 3.75 | 49.98 | 41.33 |
| WPO | -1.51 | 3.53 | 39.23 | 41.36 | -1.62 | 3.62 | 48.35 | 41.19 |
| TDPO | -1.45 | 3.71 | 41.83 | 40.95 | -1.84 | 3.61 | 46.11 | 41.14 |
| VanillaKD | -1.66 | 3.46 | 38.88 | 41.17 | -1.70 | 3.53 | 44.06 | 41.09 |
| SeqKD | -1.38 | 3.40 | 37.16 | 41.17 | -1.75 | 3.20 | 38.47 | 41.17 |
| DDPO | -1.53 | 3.61 | 37.3 | 41.23 | -1.64 | 3.58 | 43.97 | 41.23 |
| DPKD | -1.57 | 3.43 | 37.99 | 41.35 | -1.73 | 3.48 | 43.04 | 41.14 |
| GKD | -1.61 | 3.52 | 36.63 | 41.19 | -1.74 | 3.41 | 40.78 | 40.92 |
| DCKD | -1.60 | 3.55 | 36.79 | 41.33 | -1.46 | 3.51 | 49.98 | 41.09 |
| ADPA | -1.21 | 3.73 | 50.00 | 40.29 | -1.43 | 3.57 | 50.00 | 41.26 |
| **TVKD(Ours)** | **-1.15** | **3.97** | **52.18** | **41.61** | **-1.18** | **3.98** | **54.85** | **41.35** |

- Our method outperforms baselines across multiple benchmarks on two datasets.

# Experiment

| Method | DPOMIX RM(↑) | MT(↑) | AE(↑) | OLL(↑) | Helpsteer2 RM(↑) | MT(↑) | AE(↑) | OLL(↑) |
|---|---|---|---|---|---|---|---|---|
| DPO Teacher | -0.77 | 5.74 | 73.12 | 53.45 | -0.88 | 5.63 | 72.87 | 57.95 |
| SFT Teacher | -0.94 | 5.58 | 71.02 | 53.15 | -1.23 | 5.58 | 70.31 | 53.15 |
| SFT | -1.55 | 3.56 | 43.62 | 40.18 | -1.78 | 3.43 | 45.49 | 41.06 |
| DPO | -1.41 | 3.70 | 47.03 | 35.48 | -1.22 | 3.77 | 44.5 | 41.24 |
| SimPO | -1.22 | 3.65 | 46.12 | 41.5 | -1.37 | 3.75 | 49.98 | 41.33 |
| WPO | -1.51 | 3.53 | 39.23 | 41.36 | -1.62 | 3.62 | 48.35 | 41.19 |
| TDPO | -1.45 | 3.71 | 41.83 | 40.95 | -1.84 | 3.61 | 46.11 | 41.14 |
| VanillaKD | -1.66 | 3.46 | 38.88 | 41.17 | -1.70 | 3.53 | 44.06 | 41.09 |
| SeqKD | -1.38 | 3.40 | 37.16 | 41.17 | -1.75 | 3.20 | 38.47 | 41.17 |
| DDPO | -1.53 | 3.61 | 37.3 | 41.23 | -1.64 | 3.58 | 43.97 | 41.23 |
| DPKD | -1.57 | 3.43 | 37.99 | 41.35 | -1.73 | 3.48 | 43.04 | 41.14 |
| GKD | -1.61 | 3.52 | 36.63 | 41.19 | -1.74 | 3.41 | 40.78 | 40.92 |
| DCKD | -1.60 | 3.55 | 36.79 | 41.33 | -1.46 | 3.51 | 49.98 | 41.09 |
| ADPA | -1.21 | 3.73 | 50.00 | 40.29 | -1.43 | 3.57 | 50.00 | 41.26 |
| **TVKD(Ours)** | **-1.15** | **3.97** | **52.18** | **41.61** | **-1.18** | **3.98** | **54.85** | **41.35** |

- Student: LLaMA 3.2 – 1B        Teacher: LLaMA 3.2 – 8B
- Training datasets: DPOMIX and HelpSteer
- Evaluation metrics: Reward model score, MT-Bench, Alpaca Eval, and Open LLM Leaderboard

# Ablation Setting

Table 2: Results of ablation for TVKD using DPOMIX dataset. The best performances are highlighted in **bold**, while second-best performances are underline.

| | Mistral-7B->Danube-500M | | Mistral-7B -> Llama-1B | | Llama-8B->Llama-3B | |
|---|---|---|---|---|---|---|
| | RM(↑) | MT(↑) | RM(↑) | MT(↑) | RM(↑) | MT(↑) |
| DPO | -3.03 | <u>3.21</u> | -1.61 | 3.28 | <u>-1.10</u> | <u>5.08</u> |
| SimPO | -2.91 | 2.96 | <u>-1.55</u> | <u>3.32</u> | -1.12 | 4.90 |
| DCKD | <u>-2.08</u> | 2.84 | -1.72 | 3.07 | -1.25 | 4.86 |
| Ours | **-1.99** | **3.22** | **-1.36** | **3.38** | **-1.05** | **5.19** |

- The method remains robust across various teacher–student pairs.

# Ablation Setting

Table 5: Comparison of various auxiliary rewards on the same setting in Table 1. In Alpaca-Eval (AE), we use our method as a baseline.

| Type | Name | Auxiliary Reward | Margin Acc.(↑) | MT-bench(↑) | AE(↑) |
|---|---|---|---|---|---|
| Action Dependent | Logits | $Q(a, s)$ | 18.29 | 3.61 | 37.18 |
| | Log Probability | $\log \pi_\phi(a \mid s)$ | 18.31 | 3.43 | 32.50 |
| State Dependent | Max | $\max_a \pi_\phi(a \mid s)$ | 28.86 | 3.79 | 36.07 |
| | Margin | $\pi_\phi^{(1)}(s) - \pi_\phi^{(2)}(s)$ | 30.23 | <u>3.90</u> | 48.81 |
| | Expectation | $\sum_a \pi_\phi(a \mid s) Q_\phi(a \mid s)$ | 30.23 | 3.45 | <u>49.42</u> |
| | **Ours** | $\log \sum_a \exp\left(Q_\phi(a \mid s)\right)$ | 30.23 | **3.97** | **50.00** |

- We conducted ablation studies by varying the auxiliary reward term:
- Action-dependent terms significantly degrade test accuracy.
- State-dependent terms preserve performance.
- Our value-based term achieves the best performance overall.

# Thank you