

SafeDPO: A Simple Approach to Direct Preference Optimization with Enhanced Safety

Geon-Hyeong Kim, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, Youngsoo Jang, Moontae Lee

Motivation

Safety Alignment Problem

I feel upset and might do something bad.

Safe & Helpful

I can't support harm, but I can help you cope.

Safe & Unhelpful

I can't help with that

Unsafe

Act on it to feel better

Why Existing Safe Alignment Is Complex

Hard Constraint

 $c(x, y) \leq 0 \quad (\text{for all } x, y)$

→

Relaxed Constraint

 $\mathbb{E}[c(x, y)] \leq \hat{C}$

Relaxation → Auxiliary models or multi-stage optimization (e.g., SafeRLHF)

Safe RLHF

Prompt x

Response A, Response B

Reward Model r_ϕ

Cost Model c_ψ

→ PPO- λ

Hard constraint → Closed-form (intractable) → Can we make it tractable?

Ideal Optimal Policy

Safe & Helpful

Highest preference

Safe & Unhelpful

Lower preference
(feasible but suboptimal)

Unsafe

Lowest preference
(zero probability)

→ How can we realize this ordering in a tractable and principled way?

SafeDPO via Safety-Aware Transformation

SafeDPO

Safety-Aware Transformation T

Prompt x

Response A, Response B

Safe vs. Safe → Keep

Safe vs. Unsafe → Prefer safe

Unsafe vs. Unsafe → Discard

→ Standard DPO

Enforce this ordering through a safety-aware transformation of preference data.

Why SafeDPO Is Correct?

Step 1: DPO-compatible reformulation

- Hard constraint → Equivalent unconstrained objective

$$\max_{\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(\cdot | x)} [r(x, y) - \beta D_{\text{KL}}(\pi_{\theta}(\cdot | x) \| \pi_{\text{ref}}(\cdot | x))],$$

$$\text{s.t. } c(x, y) \leq 0, \quad \forall x \sim \mathcal{D}, y \sim \pi_{\theta}(\cdot | x).$$

$$\max_{\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(\cdot | x)} [r_c(x, y) - \beta D_{\text{KL}}(\pi_{\theta}(\cdot | x) \| \pi_{\text{ref}}(\cdot | x))]$$

Step 2: Tractable transformation

- Intractable objective → Equivalent tractable objective

$$\tilde{\mathcal{L}}(\theta) = -\mathbb{E}_{(x, \tilde{y}_w, \tilde{y}_l) \sim \tilde{\mathcal{D}}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(\tilde{y}_w | x)}{\pi_{\text{ref}}(\tilde{y}_w | x)} - \beta \log \frac{\pi_{\theta}(\tilde{y}_l | x)}{\pi_{\text{ref}}(\tilde{y}_l | x)} \right) \right]$$

$$\mathcal{L}_{\text{SafeDPO}}(\theta) = -\mathbb{E}_{(x, \tilde{y}_w, \tilde{y}_l) \sim T(\mathcal{D})} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(\tilde{y}_w | x)}{\pi_{\text{ref}}(\tilde{y}_w | x)} - \beta \log \frac{\pi_{\theta}(\tilde{y}_l | x)}{\pi_{\text{ref}}(\tilde{y}_l | x)} \right) \right]$$

Step 3: Safety enhancement

- Δ increases safe-unsafe separation without changing the optimal policy

$$\mathcal{L}_{\text{SafeDPO}}(\theta; \Delta) = -\mathbb{E}_{T(\mathcal{D})} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(\tilde{y}_w | x)}{\pi_{\text{ref}}(\tilde{y}_w | x)} - \beta \log \frac{\pi_{\theta}(\tilde{y}_l | x)}{\pi_{\text{ref}}(\tilde{y}_l | x)} - (\tilde{h}_l - \tilde{h}_w) \Delta \right) \right]$$

→ Safety alignment can be simplified to a DPO-style objective through a principled reformulation

Experimental Results

Overall Performance

- ★ SafeDPO
- ✦ SACPO
- SFT
- DPO-HARMLESS

- ✦ SafeRLHF
- ✦ P-SACPO
- ▲ DPO-HELPFUL
- ◆ DPO-SAFEBETTER

PKU-SafeRLHF-30K Dataset

Model-based Evaluation: Harmless_Ratio (%) vs Helpfulness

GPT-based Evaluation: Harmless_Ratio (%) vs Helpfulness

XSTest Dataset

GPT-based Evaluation: Harmless_Ratio (%) vs Over Refusal Ratio (%)

Near-perfect safety ($\approx 100\%$) → Without complex pipelines
 Strict safety → At the cost of conservativeness

Effectiveness and Sensitivity of Δ

Harmless_Ratio (%) vs Δ

Helpfulness vs Δ

Harmless_Ratio (%) vs Δ

Helpfulness vs Δ