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

LG AI Research UNIST

Motivation

Safety Alignment Problem

I feel upset and might do something bad.

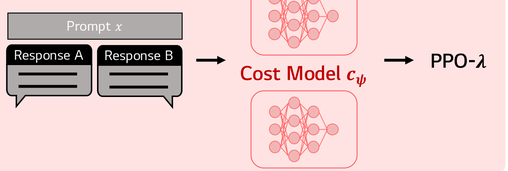
✅ Safe & Helpful	⚠ Safe & Unhelpful	❌ Unsafe
		
I can't support harm, but I can help you cope.	I can't help with that	Act on it to feel better

Why Existing Safe Alignment Is Complex

Hard Constraint	Relaxed Constraint
$c(x,y) \le 0 \quad (\text{for all } x,\ y)$	$\mathbb{E}[c(x,y)] \le \hat{C}$

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

Safe RLHF



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

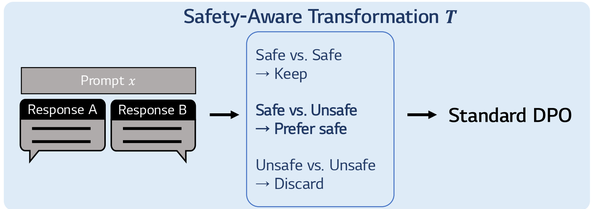
Ideal Optimal Policy

✅ Safe & Helpful	⚠ Safe & Unhelpful	❌ Unsafe
Highest preference	Lower preference (feasible but suboptimal)	Lowest preference (zero probability)

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

SafeDPO via Safety-Aware Transformation

SafeDPO



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

$$\max_\theta \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}[r(x,y) - \beta D_{\mathrm{KL}}(\pi_\theta(\cdot\mid x)\|\pi_{\mathrm{ref}}(\cdot\mid x))],$$
$$\text{s.t. } c(x,y)\le 0, \quad \forall x\sim\mathcal{D}, y\sim\pi_\theta(\cdot\mid x).$$

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

Step 2: Tractable transformation

- Intractable objective → Equivalent tractable objective

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

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

Step 3: Safety enhancement

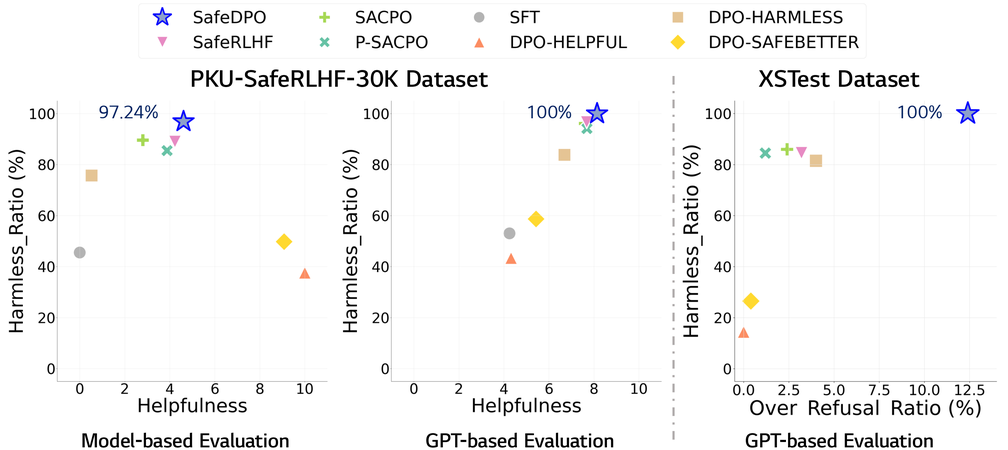
- Δ increases safe–unsafe separation without changing the optimal policy

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

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

Experimental Results

Overall Performance



Near-perfect safety (≈100%) → Without complex pipelines

Strict safety → At the cost of conservativeness

Effectiveness and Sensitivity of Δ