Stepwise Alignment for Constrained Language Model Policy Optimization

Akifumi Wachi*[1], Thien Q. Tran*[1], Rei Sato[1], Takumi Tanabe[1], Youhei Akimoto[2,3]

[1]LY Corporation (*Equal contribution) [2]University of Tsukuba [3]RIKEN AIP Correspondence: akifumi.wachi@lycorp.co.jp

NEURAL INFORMATION PROCESSING SYSTEMS

Paper Code & Models

Overview

Alignment of language models (LMs) is inherently multifaceted:

Helpfulness vs. Harmlessness

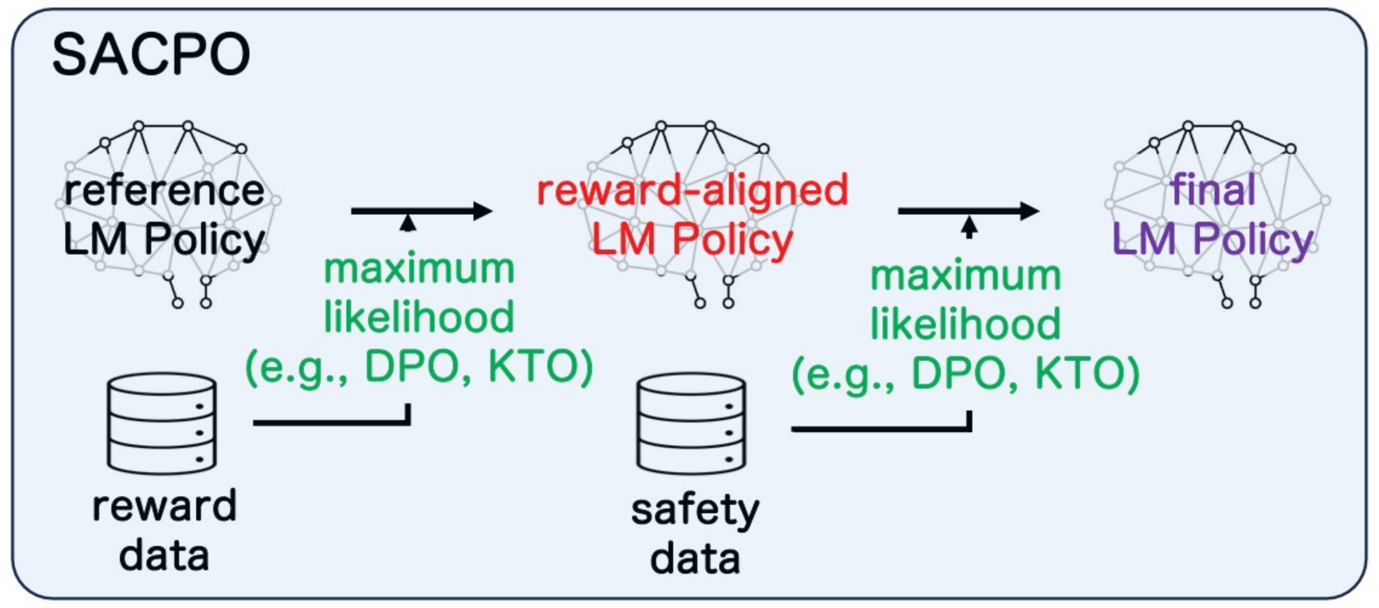
We consider the following constrained LM alignment problem:

$$\max_{\pi} \underbrace{\mathbb{E}_{\rho,\pi}[r^\star(x,y)] - \beta \mathbb{D}_{\mathrm{KL}}[\pi(y \mid x) \,\|\, \pi_{\mathrm{ref}}(y \mid x)]}_{\text{Typical objective of RLHF or DPO}} \quad \text{subject to} \quad \underbrace{\mathbb{E}_{\rho,\pi}[g^\star(x,y)] \ge b}_{\text{Safety constraint}}$$

SACPO: Stepwise Alignment for Constrained Policy Optimization

Key Idea: Reward alignment → Safety Alignment (or vice versa)

We can use RL-free alignment algorithms (e.g., DPO, KTO) for each alignment



- SACPO's stepwise approach is theoretically justified!
- SACPO is computationally efficient and stable!

Detailed Steps

Step 1: Reward Alignment

- Align an LM reference policy using reward data via an RL-free alignment algorithm (e.g., DPO, KTO)
- This step is same as typical alignment by DPO or KTO. For example,

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

Step 2: Safety Realignment

- Realign the reward-aligned LM policy using safety data using DPO or KTO
- Note : $\lambda^\star$ is the optimal Lagrangian multiplier

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta, \pi^\star_{r^\star}, \beta/\lambda^\star) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_g}\left[\log \sigma\left(\frac{\beta}{\lambda^\star} \log \frac{\pi_\theta(y_w \mid x)}{\pi^\star_{r^\star}(y_w \mid x)} - \frac{\beta}{\lambda^\star} \log \frac{\pi_\theta(y_l \mid x)}{\pi^\star_{r^\star}(y_l \mid x)}\right)\right]$$

	Step1 (reward alignment)	Step 2 (safety realignment)
Loss function	$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta, \pi_{\mathrm{ref}}, \beta)$	$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta, \pi^\star_{r^\star}, \beta/\lambda^\star)$
LM policy to be aligned	π_{ref} (typically SFT models)	$\pi_{r^\star}$ (reward-aligned LM)
KL penalty parameter	β	$\beta/\lambda^\star$

Why is SACPO Theoretically Justified?

The optimal policy of our constrained LM alignment problem satisfies:

$$\boxed{\pi^\star(y \mid x)} \propto \boxed{\pi_{\mathrm{ref}}(y \mid x) \exp\left(\frac{1}{\beta}(r^\star(x,y)\right.} + \lambda^\star \boxed{g^\star(x,y)}\Big)\Big)$$

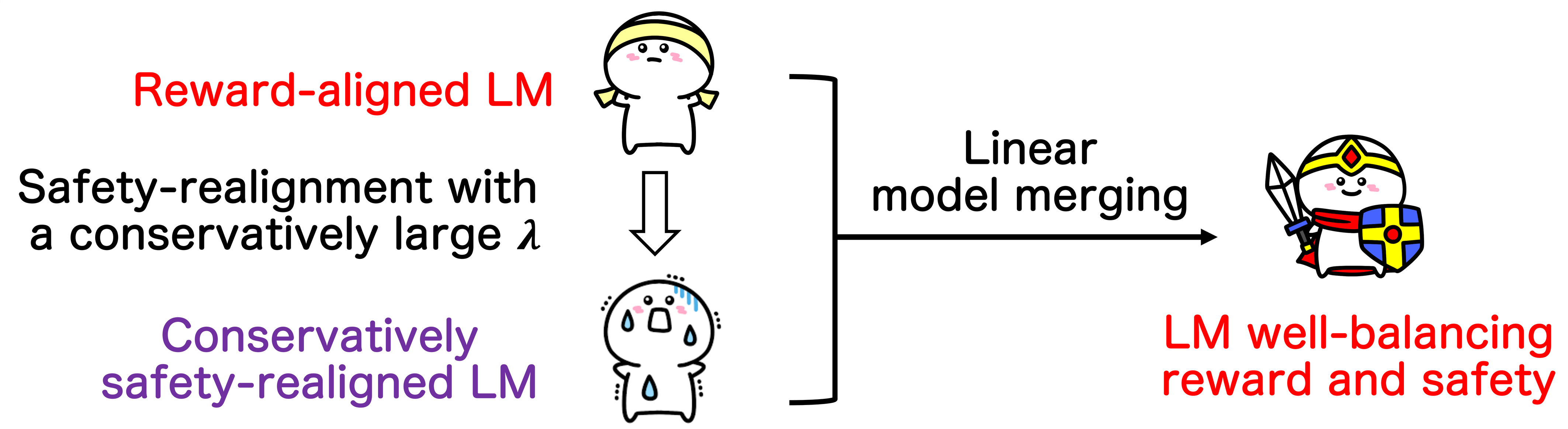
$$\propto \underbrace{\boxed{\pi^\star_{r^\star}(y \mid x)}}_{\text{Reward-aligned LM policy}} \exp\left(\frac{\lambda^\star}{\beta} \underbrace{\boxed{g^\star(x,y)}}_{\text{Safety function}}\right)$$

The optimal policy can be obtained by realigning reward-aligned LM policy regarding the safety function with a KL penalty parameter $\beta/\lambda^\star$.

Practical SACPO (P-SACPO)

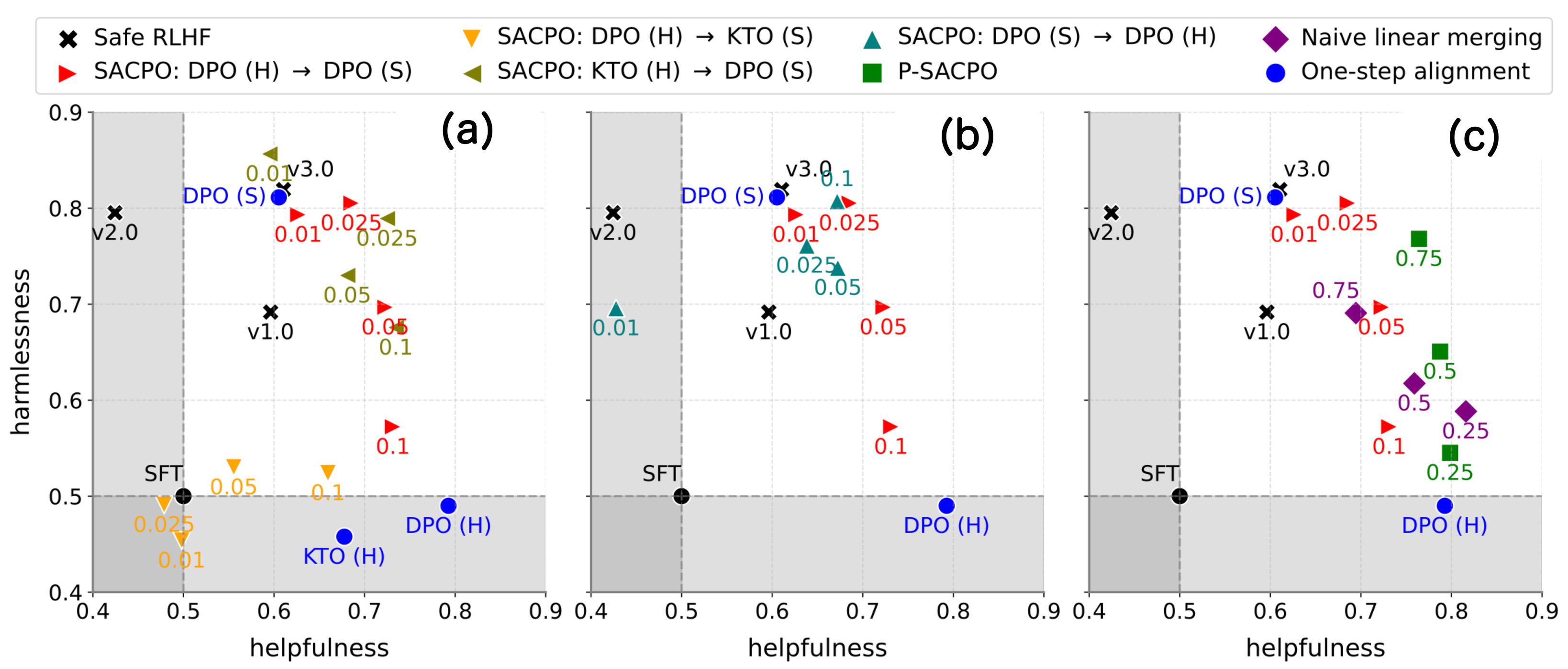
It is still costly to apply DPO (or KTO) w/ various $\beta/\lambda^\star$ in SACPO

Control the balance between reward and safety by tuning the merging ratio!



Experiments

- Compare the performance of **(P-)SACPO** with **Safe RLHF**
- Reward = Helpfulness (H), Safety = Harmlessness (S)
- SFT model = Alpaca-7b, Dataset = PKU-SafeRLHF-30K



Win-rate against SFT model. In (a) and (b), the numbers indicate β/λ. In (c), the numbers for the red triangles represent β/λ, while those for the green and purple squares represent the model merging ratio.

Results

1. SACPO well-balances helpfulness and safety, which performs better than Safe RLHF!
2. In SACPO, difference alignment algorithms can be used (this is consistent with theory)
3. P-SACPO empirically performs well with reduced computational time and stable learning