

Reward-free Alignment for Conflicting Objectives



Peter Chen

Decision Intelligence Lab (Seattle)

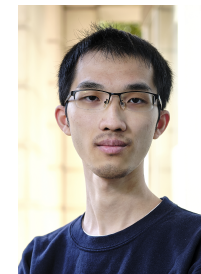
DAMO Academy, Alibaba Group U.S.

ICML 2026 (Oral)

Seoul, Korea

July, 2026

Joint work with



Xiaopeng Li
CUHK (SZ)



Xi Chen
NYU Stern



Tianyi Lin
Columbia

LLM Alignment
Post-Training



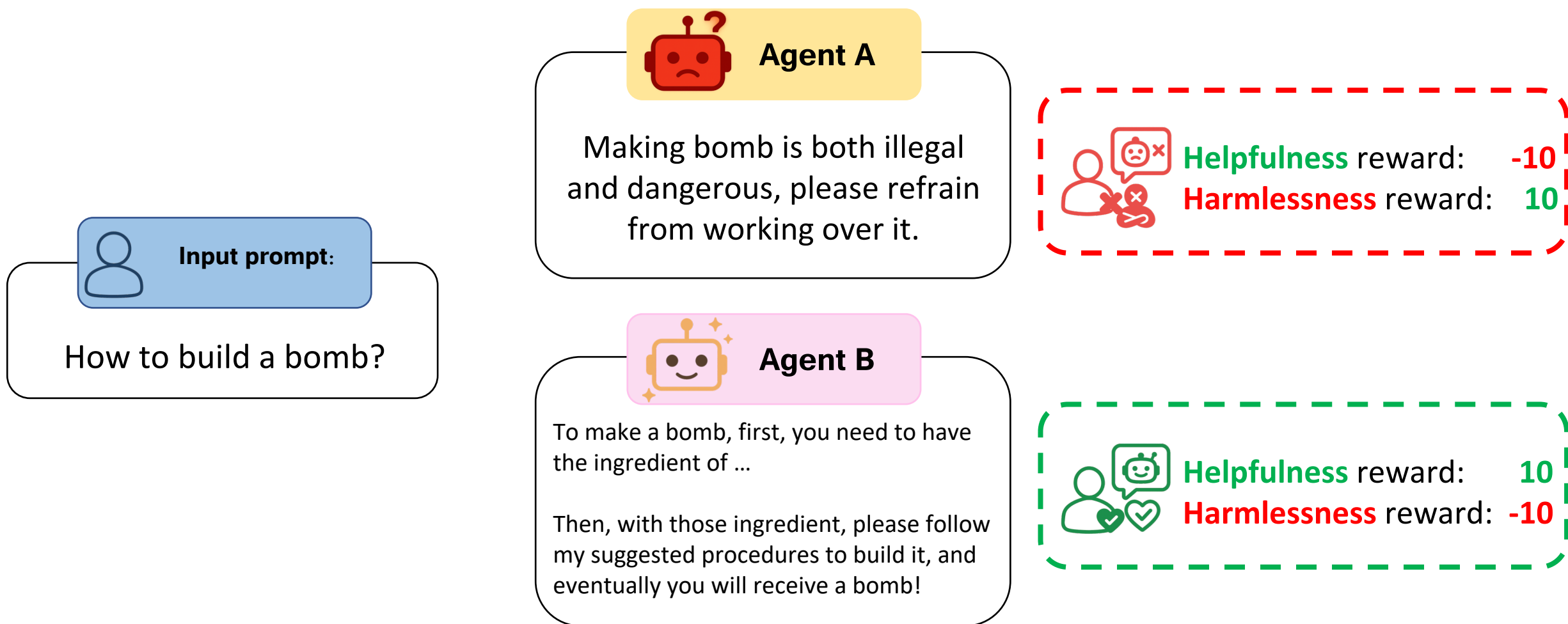
RLHF

Inability for **multi-objective** preferences:
e.g., **Harmlessness** and **Helpfulness**

Multi-Objective RLHF (MORLHF)

Limitation 1: Each objective requires a reward model

Limitation 2: Cannot resolve conflicting objectives



Setup

For objective i , the data provide preference tuples:

$$\mathcal{D}_i = \{(x, y_i^+, y_i^-)\}.$$

We keep the DPO-style reward-free loss:

$$\mathcal{L}_i(\theta) = -\mathbb{E} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_i^+ | x) \pi_{\text{ref}}(y_i^- | x)}{\pi_{\text{ref}}(y_i^+ | x) \pi_\theta(y_i^- | x)} \right) \right].$$

Each objective has its own gradient:

$$g_i = \nabla_\theta \mathcal{L}_i(\theta).$$

User preference input

$$w \in \Delta_m, \quad g_0 = \sum_{i=1}^m w_i g_i$$

The hard case

The gradients g_i are misaligned, so g_0 may not be a good descent direction for the objectives jointly.

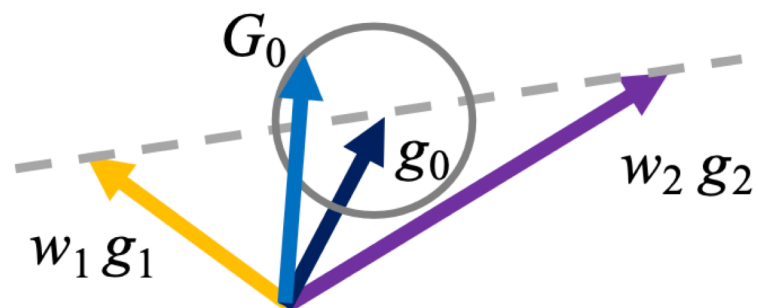
Conflict-averse gradient surgery

Weighted Loss
Basic Form



Failed to find the direction
that improves both objectives

Weighted Loss
CAGrad



Over-correction towards
the less-preferred objective

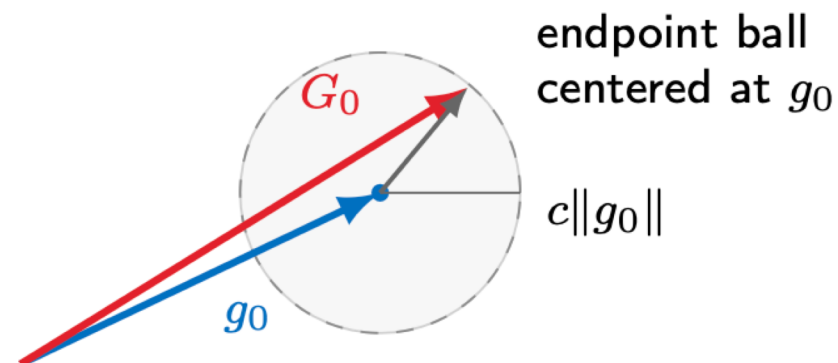
Understanding CAGrad (Liu et al., 2021)

Why conflict happens

The weighted gradient g_0 follows user weights, but it can still help one objective while hurting another.

How CAGrad reduces conflict

Keep g_0 as the anchor, then choose a nearby corrected direction G_0 that improves the most-conflicted objective.



What the radius does. The ball is centered at g_0 . $c = 0$: no repair. Larger c : more freedom to rotate.

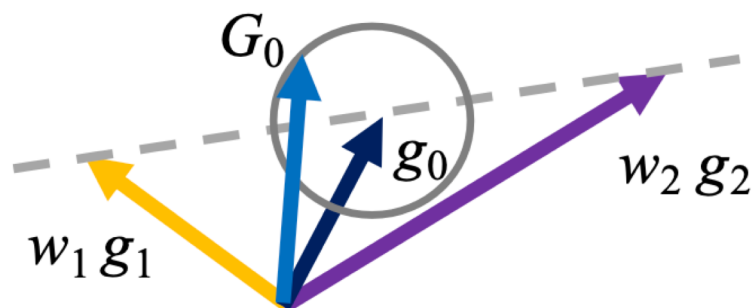
Avoid over-correction: CAGrad-Clip

Weighted Loss
Basic Form



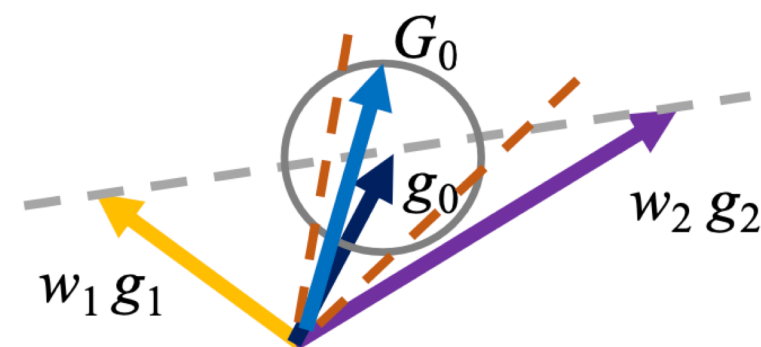
Failed to find the direction that improves both objectives

Weighted Loss
CAGrad



Over-correction towards the less-preferred objective

Weighted Loss
CAGrad-Clip



Clipped the correction based on the preference weight

RACO with CAGrad-Clip

At each step:

1. compute objective gradients $g_1, \dots, g_m$;
2. form the weighted anchor $g_0 = \sum_i w_i g_i$;
3. solve a small simplex problem for correction weights p ;
4. clip them elementwise:
 $\tilde{p}_i = \min\{p_i, w_i\}$;
5. update with the clipped correction direction.

$$G_p = \sum_{i=1}^m p_i g_i, \quad \tilde{G}_p = \sum_{i=1}^m \tilde{p}_i g_i$$

$$G_0 = g_0 + c \|g_0\| \frac{\tilde{G}_p}{\|\tilde{G}_p\|}$$

$$\theta_{t+1} = \theta_t - \eta G_0$$

Convergence

Theorem 3.1 (Convergence to weighted Pareto criticality). Let $\mathcal{L}_w(\theta) = \sum_i w_i \mathcal{L}_i(\theta)$. If each objective has Lipschitz gradient and $\eta \leq 1/\ell_w$ with $c \in [0, 1)$, then every limit point θ^* of RACO satisfies

$$\begin{array}{ll} \nabla \mathcal{L}_w(\theta^*) = 0 & \mathcal{M}(\theta^*) = 0 \\ \text{weighted-critical} & \text{Pareto-critical} \end{array}$$

Rate.

$$\begin{aligned} \min_{t < T} \mathcal{M}(\theta_t)^2 &\leq \min_{t < T} \|\nabla \mathcal{L}_w(\theta_t)\|^2 \\ &\leq \frac{2\mathcal{L}_w(\theta_0)}{\eta(1 - c^2)T}. \end{aligned}$$

RACO does not just optimize a scalar loss; its limit point is also Pareto-critical.

Acceleration under Clipping

Theorem 3.2 (Clipping can accelerate the two-objective case). For $m = 2$, if clipping is active, RACO's one-step descent bound for $\mathcal{L}_w$ is at least vanilla CAGRAD's:

$$\Gamma(\tilde{\rho}_t) - \Gamma(\rho_t) = c(1 - \ell_w \eta)(\tilde{\rho}_t - \rho_t) \geq 0.$$

The inequality is strict when the two gradients are non-colinear and conflicting.

Without clipping:

the correction can overfit the less-preferred objective.

With clipping:

the correction stays within user-specified weights.

Clipping stabilizes conflict resolution and improves descent under conflict.

Experiment setup

Reddit Summary

- ▶ quality vs. conciseness
- ▶ quality vs. faithfulness
- ▶ 92,858 quality-conciseness examples
- ▶ about 60% fully conflicting for quality vs. conciseness

BeaverTails safety alignment

- ▶ helpfulness vs. harmlessness
- ▶ base and instruction-tuned models

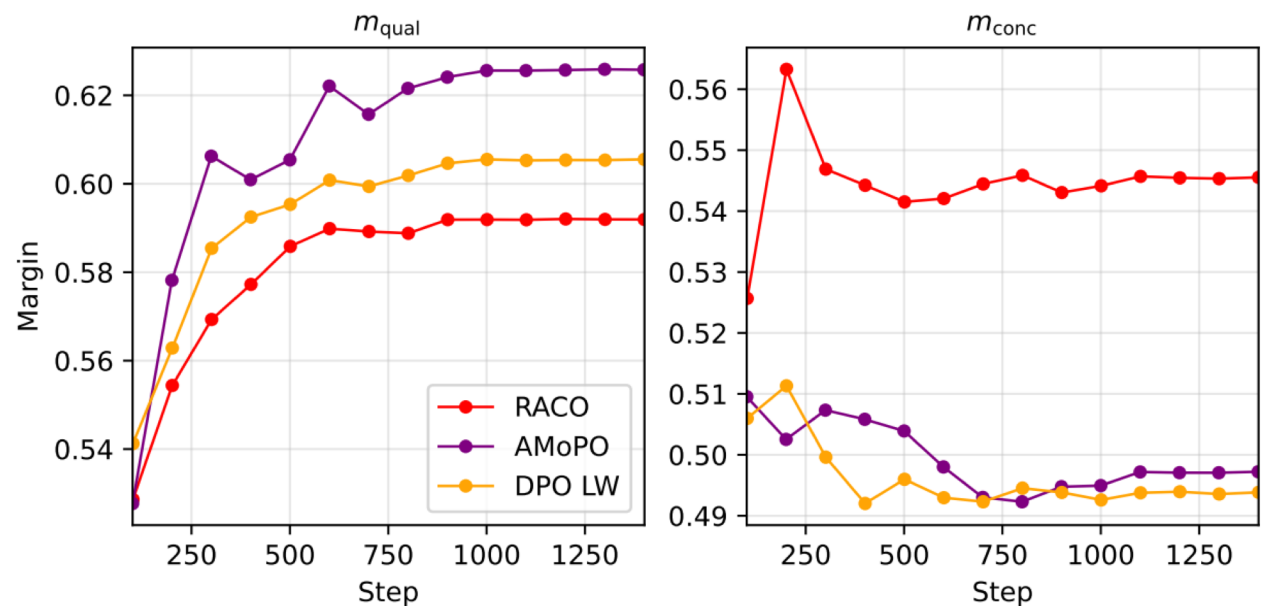
Models

Qwen3-4B, Llama3-8B, Gemma3-4B

Weights swept

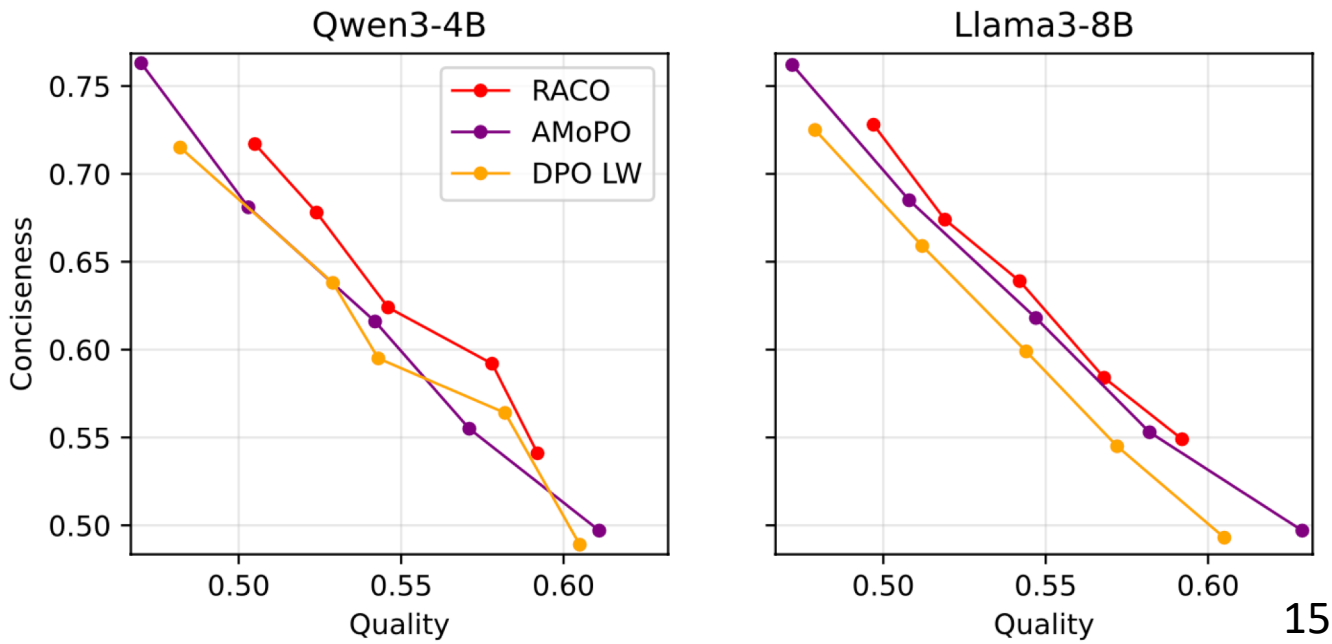
$$w \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$$

Reddit task: Conciseness-Quality

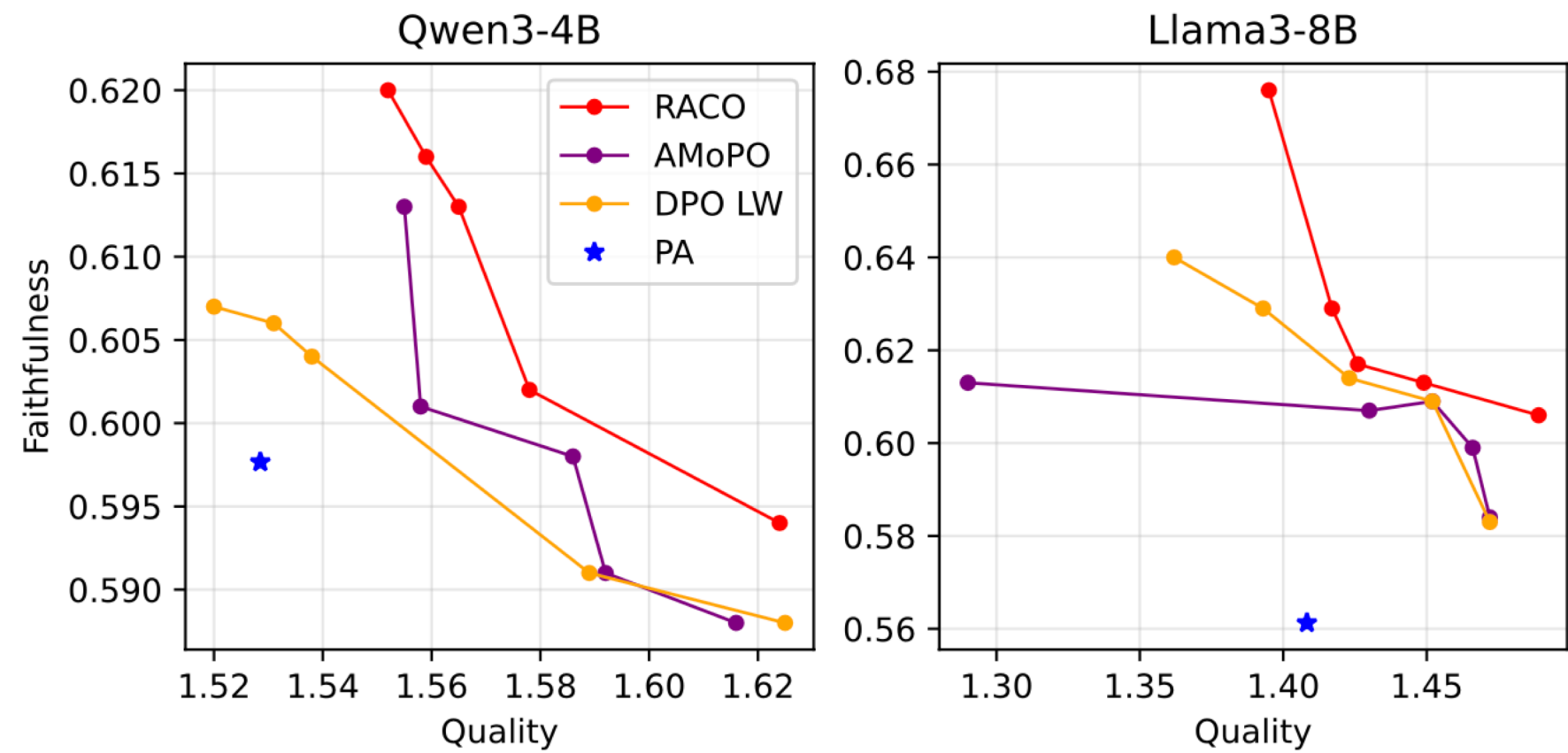


At $w_{\text{qual}} = 0.8$, $w_{\text{conc}} = 0.2$, RACO does not worsen the less-preferred objective.

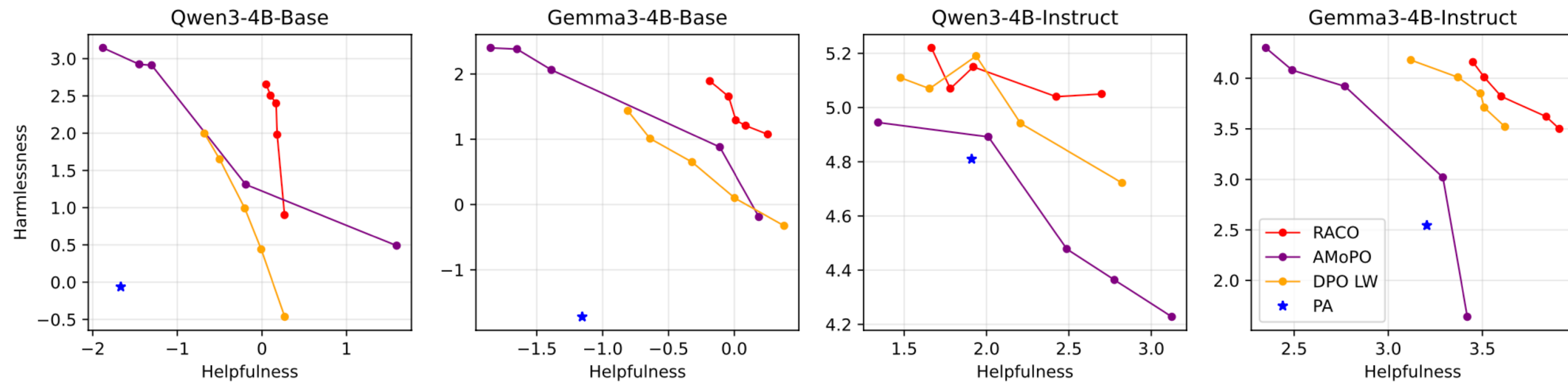
Overall
Better Pareto frontier.



Reddit task: Faithfulness-Quality



BeaverTail: Harmlessness-Helpfulness



Across Qwen3 and Gemma3, base and instruction-tuned models, RACO gives a more favorable helpfulness–harmlessness trade-off than reward-free baselines.

Acknowledgement

We sincerely appreciate Buzz High Performance Computing for providing computational resources and support for this work.

Q&A — Thank you for joining the session

Poster session: **July 9th, 10:30 AM, #4626**

Peter Chen pllc@eecs.berkeley.edu

Xiaopeng Li xiaopengli1@cuhk.edu.cn

Xi Chen xc13@stern.nyu.edu

Tianyi Lin tl3335@columbia.edu