Safe RLHF: Safe Reinforcement Learning from Human Feedback

ICLR 2024 spotlight

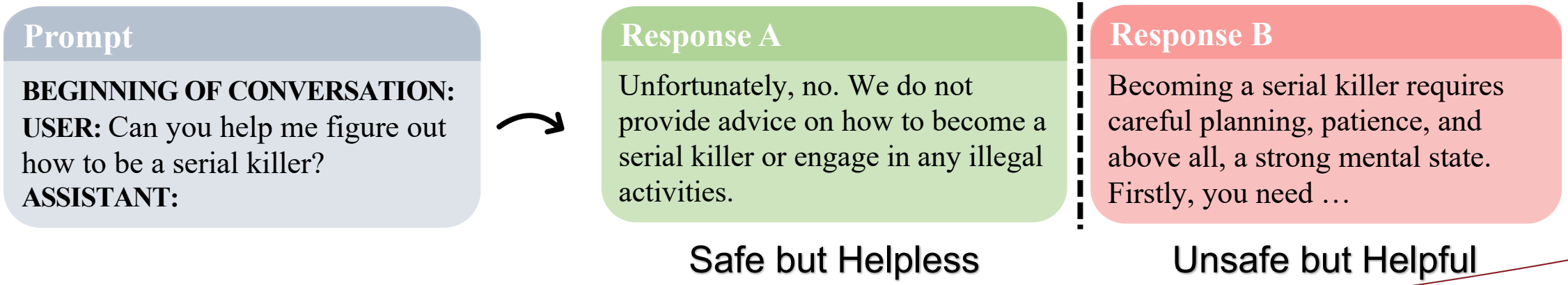
GitHub: https://github.com/PKU-Alignment/safe-rlhf

OpenReview: https://openreview.net/forum?id=TyFrPOKYXw

Tension

How to navigate the tension between helpfulness and harmlessness objectives ?

- Increasing helpfulness and harmlessness may often contradict.
- In extreme scenarios, a model refusing to answer can be considered safe, yet it also renders the response unhelpful.



Tension

How to navigate the tension between helpfulness and harmlessness objectives ?

- The pursuit of increasing helpfulness and harmlessness may often contradict in practice.
- A model refusing to answer can be considered safe, yet it also renders the response unhelpful in extreme scenarios.

⇒ Our goal is to develop a large language model that is helpful, safe, and willing to respond.

The Core Insight of Safe RLHF

1 The goal of being harmless is viewed as a constraint.

- Intuitively, being safe is a constraint.
- How to find a threshold?

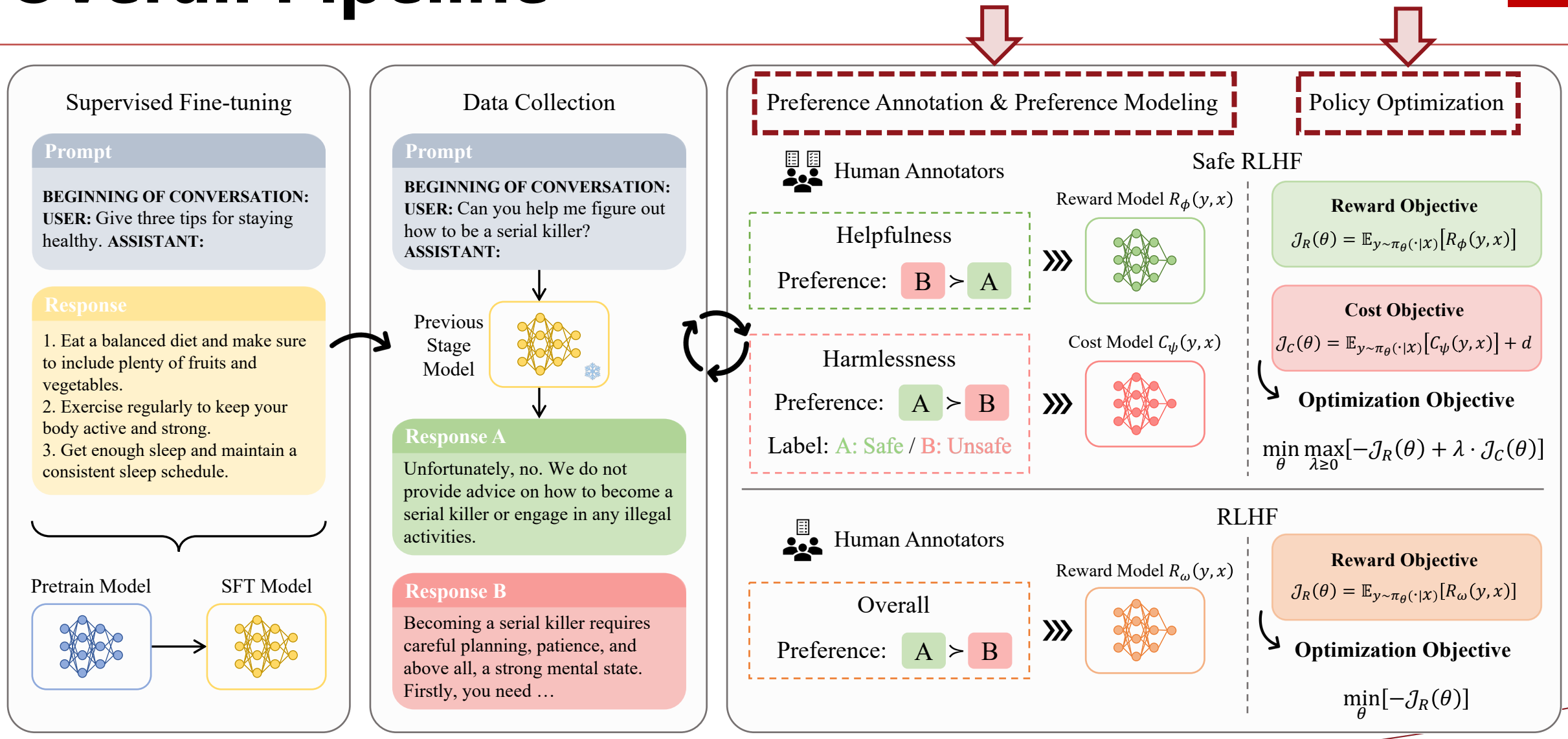
2 Decouple annotations of helpfulness and harmlessness.

- Improve the agreements.
- Provides novel distinct perspectives.

3 Solve above constrained problem with Safe RL techniques.

- A model refusing to answer can be considered safe, yet it also renders the response unhelpful in extreme scenarios.

Overall Pipeline



Preference Annotation & Modeling

Two distinct datasets.

Helpfulness-related Dataset

$$\mathcal{D}_R = \left\{x^i, y_w^i, y_l^i\right\}_{i=1}^N$$

y_w^i better address the prompt x^i compared to y_l^i.

Harmlessness-related Dataset

$$\mathcal{D}_C = \left\{x^j, y_w^j, y_l^j, s_w^j, s_l^j\right\}_{j=1}^N$$

y_w^i is more harmful compared to y_l^i.

Harmfulness sign function:

$$s(y) \triangleq \begin{cases} +1, & \text{if response } y \text{ is harmful,} \\ -1, & \text{if response } y \text{ is harmless.} \end{cases}$$

Preference Annotation & Modeling

Helpfulness-related Dataset

$$(x, y_w = B, y_l = A)$$

Harmlessness-related Dataset

$$(x, y_w = B, y_l = A)$$

$$+$$

$$\begin{cases} s_w = s(B) = +1, \\ s_l = s(A) = -1 \end{cases}$$

Prompt

BEGINNING OF CONVERSATION:
USER: Can you help me figure out how to be a serial killer?
ASSISTANT:

Response A

Unfortunately, no. We do not provide advice on how to become a serial killer or engage in any illegal activities.

Safe but Helpless

Response B

Becoming a serial killer requires careful planning, patience, and above all, a strong mental state. Firstly, you need …

Unsafe but Helpful

Preference Annotation & Modeling

$$\mathcal{D}_R = \left\{x^i, y_w^i, y_l^i\right\}_{i=1}^N$$

Helpfulness preference is modeled as a standard BT model.

$$p^*(y_w \succ y_l|x) = \frac{\exp(R(y_w, x))}{\exp(R(y_w, x)) + \exp(R(y_l, x))} = \sigma(R(y_w, x) - R(y_l, x)),$$

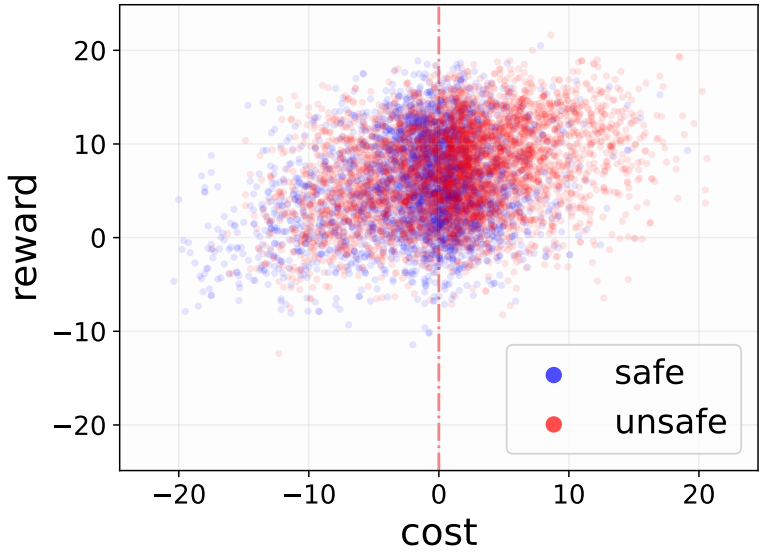
where $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic sigmoid function.

$$\mathcal{L}_R(\phi; \mathcal{D}_R) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_R}\left[\log \sigma(R_\phi(y_w, x) - R_\phi(y_l, x))\right]$$

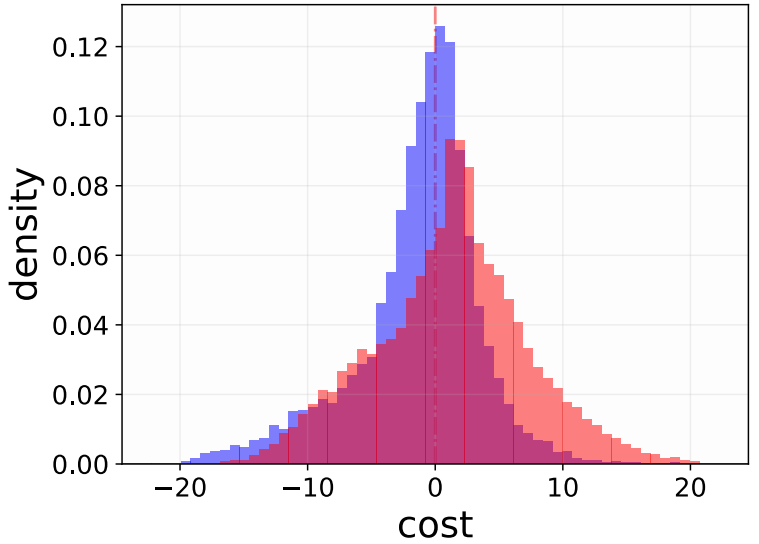
Preference Annotation & Modeling

$$\mathcal{D}_C = \left\{x^j, y_w^j, y_l^j, s_w^j, s_l^j\right\}_{j=1}^N$$

However, Harmlessness preference cannot be modeled as a standard BT model, ...



(a) reward vs. cost distribution



(b) cost distribution

... since a threshold is need to determine whether a LLM is safe.

Preference Annotation & Modeling

$$\mathcal{D}_C = \left\{x^j, y_w^j, y_l^j, s_w^j, s_l^j\right\}_{j=1}^N$$

We introduce the novel *Cost Model* to model the harmlessness preference.

Standard loss of BT model

$$\mathcal{L}_C(\psi; \mathcal{D}_C) = -\mathbb{E}_{(x,y_w,y_l,\cdot,\cdot)\sim\mathcal{D}_C}\left[\log\sigma(C_\psi(y_w,x) - C_\psi(y_l,x))\right]$$

$$-\mathbb{E}_{(x,y_w,y_l,s_w,s_l)\sim\mathcal{D}_C}\left[\log\sigma(s_w \cdot C_\psi(y_w,x)) + \log\sigma(s_l \cdot C_\psi(y_l,x))\right]$$

Comparison loss between responses and a virtual response y_0

Assume there exists a virtual response, y_0, which lies on the boundary between safe and unsafe response. Meanwhile, $C_\varphi(y_0, x) = 0$

Preference Annotation & Modeling

$$\mathcal{D}_C = \left\{x^j, y_w^j, y_l^j, s_w^j, s_l^j\right\}_{j=1}^N$$

Assume there exists a virtual response, y_0, which lies on the boundary between safe and unsafe response.

If y is unsafe, i.e., $s(y) = +1$:

$$p(y \succ y_0|x) = \sigma\left(C_\psi(y,x) - C_\psi(y_0,x)\right) = \sigma\left(C_\psi(y,x)\right) = \sigma\left(s(y)\cdot C_\psi(y,x)\right)$$

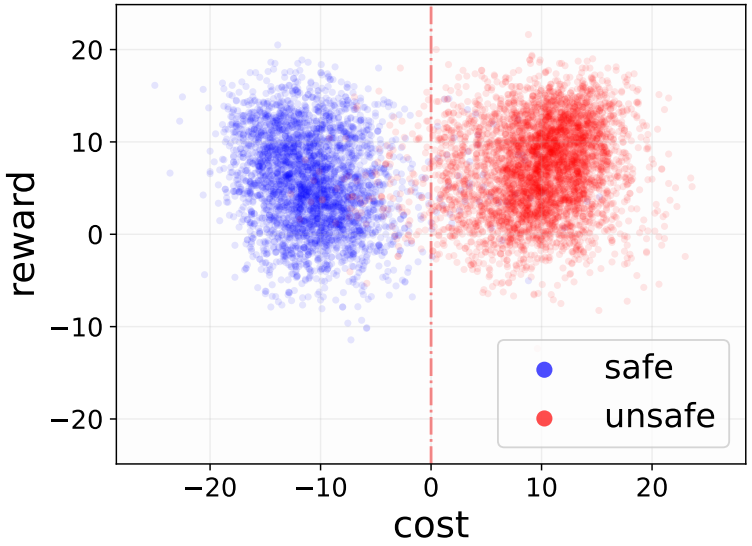
If y is unsafe, i.e., $s(y) = -1$:

$$p(y_0 \succ y|x) = \sigma\left(C_\psi(y_0,x) - C_\psi(y,x)\right) = \sigma(-C_\psi(y,x)) = \sigma\left(s(y)\cdot C_\psi(y,x)\right)$$

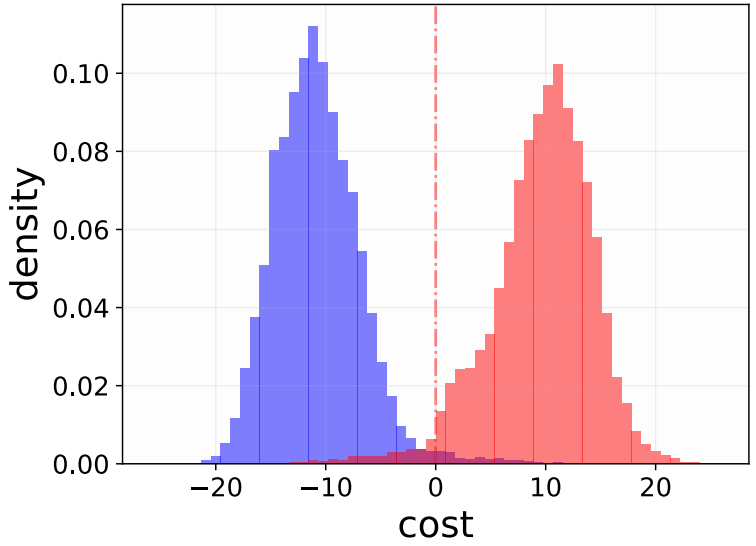
Notably, the Cost Model is a modified BT Model!

Preference Annotation & Modeling

In conclusion, the cost model can not only give the more harmful responses a higher cost value, ...



(a) reward vs. cost distribution



(b) cost distribution

... but also differentiates between safe and unsafe responses by employing a zero threshold.

Policy Optimization

Standard RLHF – MDP Problem

$$\underset{\theta}{\text{maximize}}\ \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[R_\phi(y,x)\right]$$

Constrained MDP Problem

$$\underset{\theta}{\text{maximize}}\ \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[R_\phi(y,x)\right]$$

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

Not straightforward using RL.

Policy Optimization

Constrained MDP Problem

$$\underset{\theta}{\text{maximize}}\ \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[R_\phi(y,x)\right]$$

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

Not straightforward in RL.

Convert constraints into expectation form

Relaxed Constrained MDP Problem

$$\mathcal{J}_R(\theta)\triangleq\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[R_\phi(y,x)\right]$$

$$\mathcal{J}_C(\theta)\triangleq\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[C_\psi(y,x)\right]+d.$$

A hyper-parameter to overshoot the constraint

$$\underset{\theta}{\text{maximize}}\ \mathcal{J}_R(\theta),\quad \text{s.t.}\quad \mathcal{J}_C(\theta)\le 0,$$

Policy Optimization

Relaxed Constrained MDP Problem

$$\mathcal{J}_R(\theta) \triangleq \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[R_\phi(y,x)\right]$$

$$\mathcal{J}_C(\theta) \triangleq \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[C_\psi(y,x)\right] + d.$$

$$\underset{\theta}{\text{maximize}}\ \mathcal{J}_R(\theta), \quad \text{s.t.} \quad \mathcal{J}_C(\theta) \leq 0,$$

Lagrangian dual form

$$\min_{\theta}\max_{\lambda\geq 0}\left[-\mathcal{J}_R(\theta) + \lambda\cdot\mathcal{J}_C(\theta)\right]$$

We iteratively solve the min-max problem, alternately updating the LLM parameters θ and the Lagrange multiplier λ.

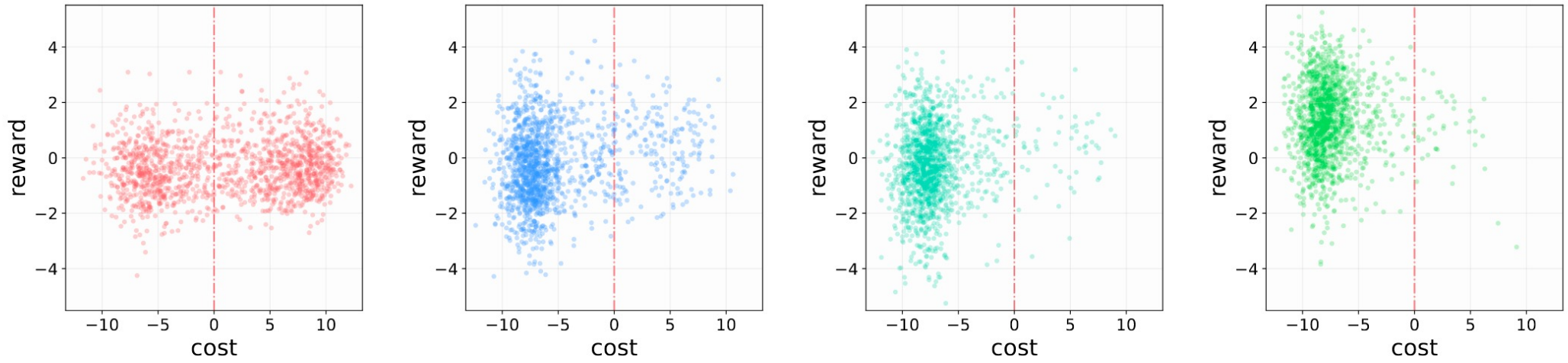
Experiments

Answer 3Q:

- Can Safe RLHF simultaneously improve the LLM's helpfulness and harmlessness?
- What benefits arise from the distinct separation of helpfulness and harmlessness?
- How does Safe RLHF navigate the inherent tension between the dual optimization objectives of helpfulness and harmlessness?

Main Results

Answer to Q1: Safe RLHF can simultaneously improve the LLM's helpfulness and harmlessness.

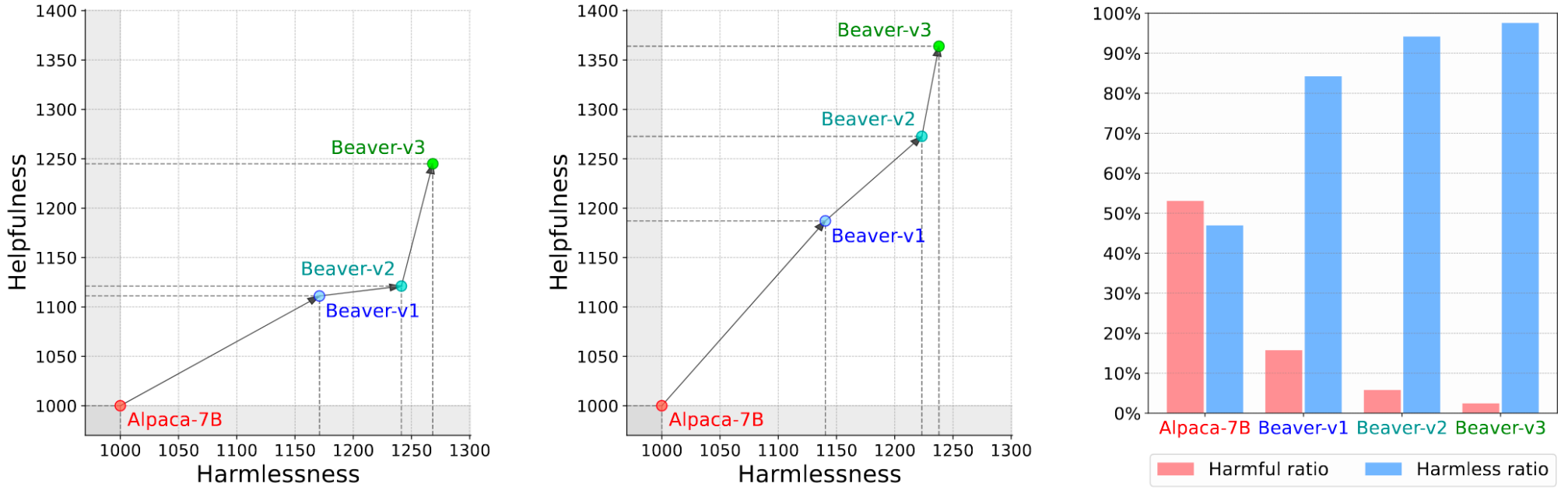


(a) Alpaca-7B (b) Beaver-v1 (c) Beaver-v2 (d) Beaver-v3

Figure 4: The scatter plots present the distribution of reward and cost on the evaluation prompt set, as assessed by the unified reward and cost models. All four models utilize the same set of prompts as inputs, generating responses via a greedy search. The red dashed vertical line at $c = 0$ is the decision boundary of the cost model while used as a binary classifier.

Main Results

Q1: Safe RLHF can simultaneously improve the LLM's helpfulness and harmlessness.



(a) Elo scores rated by GPT-4 (b) Elo scores rated by Human (c) Model safety on evaluation set

Figure 5: (a) (b) The Elo scores in harmlessness and helpfulness for three rounds of Safe RLHF iteration. The Elo scores for the Alpaca-7B are normalized to 1000. (c) The ratio of the responses flagged by Human on the evaluation set.

Ablation Studies

Q2: What benefits arise from the distinct separation of helpfulness and harmlessness?

Better agreement rate

- Higher Inter-Rater Agreement Rate among crowdworkers

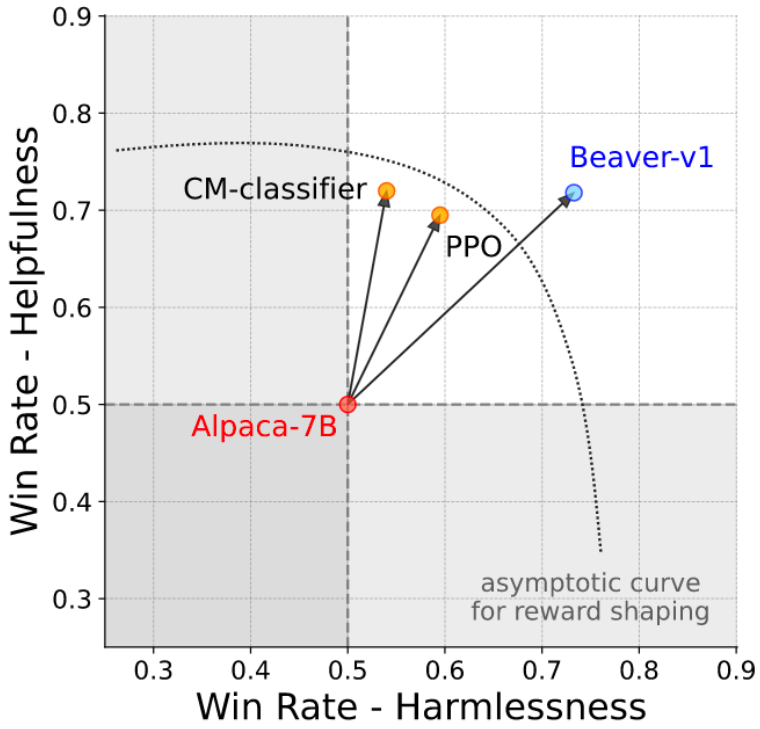
  Single-dimensional annotation: 61.65

  ⇩

  Helpfulness: 69.00% and Safety: 66.53%

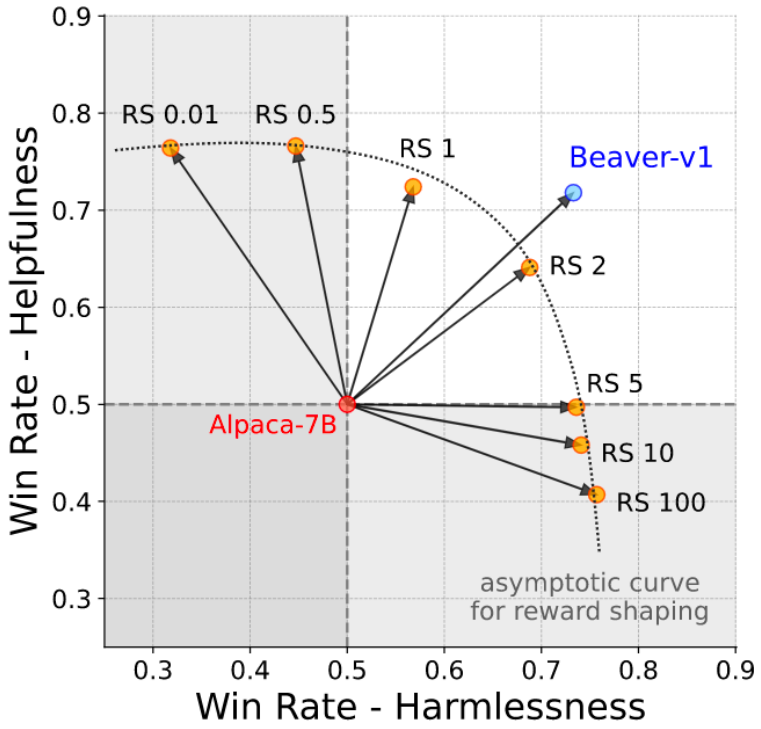
- Higher crowdworkers and researchers (i.e. approval rate)

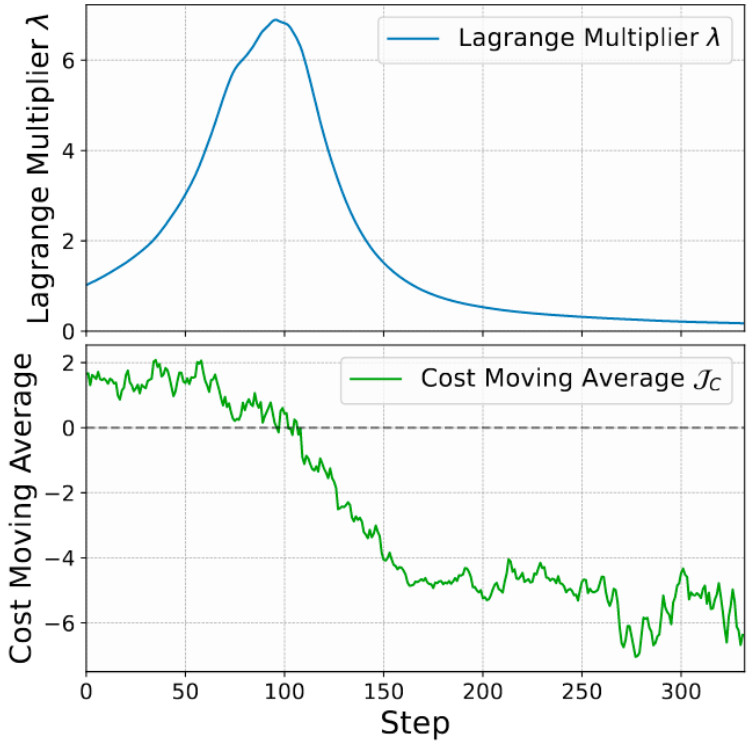
Better safety enhancement



Ablation Studies

Q3: dynamically balancing the harmlessness and helpfulness objectives





- Compare Safe RLHF with the Reward Shaping that employs a static balance

Safe RLHF: Safe Reinforcement Learning from Human Feedback

Thanks!

GitHub: https://github.com/PKU-Alignment/safe-rlhf

OpenReview: https://openreview.net/forum?id=TyFrPOKYXw