One-Shot Safety Alignment for Large Language Models via Optimal Dualization

Xinmeng Huang*† Shuo Li*† Edgar Dobriban
Osbert Bastani Hamed Hassani Dongsheng Ding†

University of Pennsylvania

Motivation

Safety requirements for language models (LM)

1. MUST NOT contain offensive or discriminatory content
2. MUST NOT fabricate content and spread misinformation

...

Constraining LMs w/ safety requirements

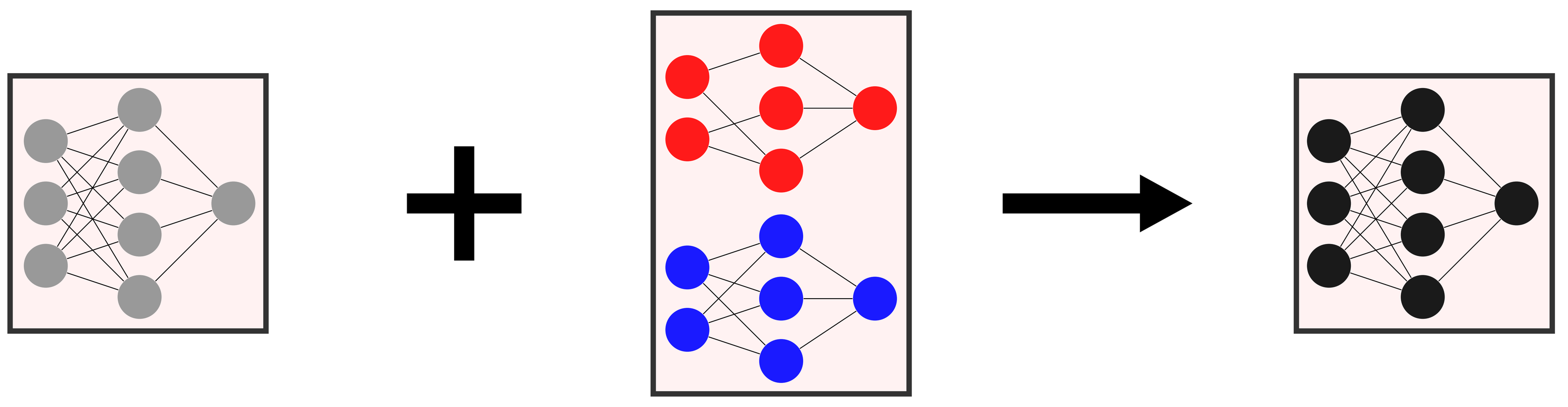
- Safe RLHF (ICLR 2024)
- Constrained RLHF (ICLR 2024)
- Constrained DPO (arXiv 2023)
- SACPO (arXiv 2024)

Hurdle • instability of iterative training • no optimality certificate

Can we align LMs w/ safety constraints in a one-shot way?

Problem formulation

LM alignment via human feedback



SFT LM + reward, safety utility models (human preferences) → aligned LM

- $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$ – (prompt, response)
- $\pi_{\text{ref}}(\cdot \,|\, \boldsymbol{x}) \in \Delta(\mathcal{Y})$ – SFT LM
- $r(\boldsymbol{x}, \boldsymbol{y})$, $g_j(\boldsymbol{x}, \boldsymbol{y})$, $j = 1, \ldots, m$ – reward, safety utility models

Constrained alignment problem

$$\max_{\pi \in \Pi} \; \mathbb{E}_{\boldsymbol{x}} \left[\mathbb{E}_{\boldsymbol{y} \sim \pi(\cdot \,|\, \boldsymbol{x})} [r(\boldsymbol{x}, \boldsymbol{y})] \; - \; \beta \, \mathrm{KL}(\pi(\cdot \,|\, \boldsymbol{x}) \| \pi_{\text{ref}}(\cdot \,|\, \boldsymbol{x})) \right]$$

$$\text{s.t.} \;\; \mathbb{E}_{\boldsymbol{x}} \left[\mathbb{E}_{\boldsymbol{y} \sim \pi(\cdot \,|\, \boldsymbol{x})} [g_j(\boldsymbol{x}, \boldsymbol{y})] \; - \; \mathbb{E}_{\boldsymbol{y} \sim \pi_{\text{ref}}(\cdot \,|\, \boldsymbol{x})} [g_j(\boldsymbol{x}, \boldsymbol{y})] \right] \; \geq \; b_j$$

- Π – LM policy set • b_j – safety margin $\quad j = 1, \ldots, m$

Optimal dualization

Dual problem

$$\underset{\boldsymbol{\lambda} \geq 0}{\text{minimize}} \quad D(\boldsymbol{\lambda}) := \underset{\pi \in \Pi}{\text{maximize}} \; L(\pi, \boldsymbol{\lambda})$$

- $L(\pi, \boldsymbol{\lambda})$ – Lagrangian

$$= \; \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{y} \sim \pi(\cdot \,|\, \boldsymbol{x})} [r(\boldsymbol{x}, \boldsymbol{y})] - \beta \mathrm{KL}(\pi(\cdot \,|\, \boldsymbol{x}) \,|\, \pi_{\text{ref}}(\cdot \,|\, \boldsymbol{x})) \quad \text{objective}$$

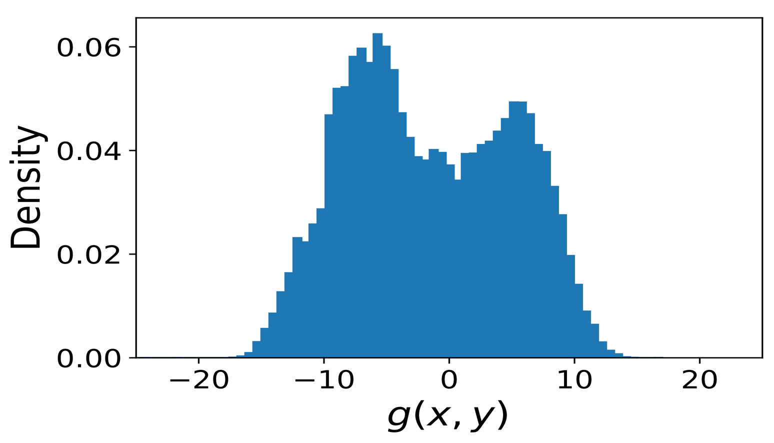
$$+ \sum_{j=1}^{m} \lambda_j \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{y} \sim \pi(\cdot \,|\, \boldsymbol{x})} [h_j(\boldsymbol{x}, \boldsymbol{y})] \quad \text{safety violation}$$

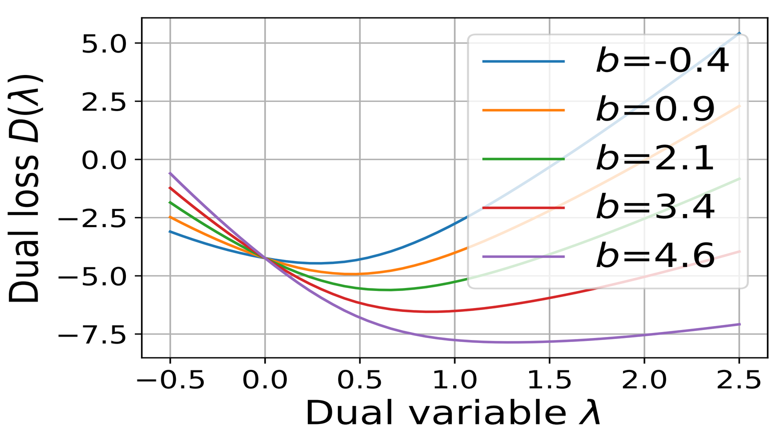
- $h_j(\boldsymbol{x}, \boldsymbol{y}) := g_j(\boldsymbol{x}, \boldsymbol{y}) - \mathbb{E}_{\pi_{\text{ref}}} [g_j(\boldsymbol{x}, \boldsymbol{y})] - b_j$ – shifted safety utility
- $L(\pi^\star, 0) = D(\boldsymbol{\lambda}^\star)$ for an optimal pair $(\pi^\star, \boldsymbol{\lambda}^\star)$ – strong duality

Explicit dual function

$$D(\boldsymbol{\lambda}) \; = \; \beta \, \mathbb{E}_{\boldsymbol{x}} \left[\ln \mathbb{E}_{\boldsymbol{y} \sim \pi_{\text{ref}}(\cdot \,|\, \boldsymbol{x})} \left[\exp \left(\frac{r(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{\lambda}^\top \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y})}{\beta} \right) \right] \right]$$

- $(\pi^\star, \boldsymbol{\lambda}^\star)$ – uniqueness of optimal primal-dual pair
- D – smooth & strongly convex at the unique $\boldsymbol{\lambda}^\star$





One-Shot Safety Alignment

Constrained Alignment via dualizatioN (CAN)

Stage 1 Optimal dual: $\boldsymbol{\lambda}^\star \; = \; \underset{\boldsymbol{\lambda} \geq 0}{\arg\min} \; D(\boldsymbol{\lambda})$

Stage 2 Update LM: $\pi^\star \; = \; \underset{\pi \in \Pi}{\arg\max} \; L(\pi, \boldsymbol{\lambda}^\star)$

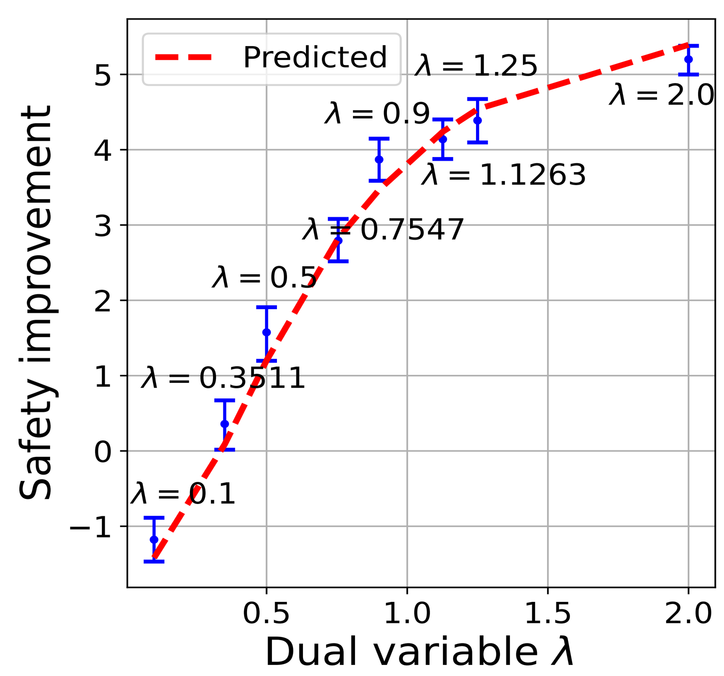
smooth convex optimization & unconstrained alignment

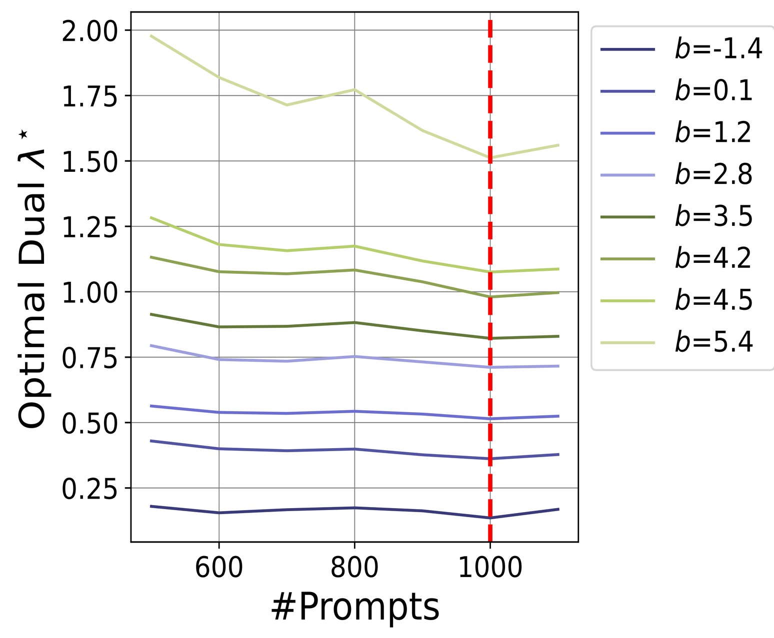
Advantages • optimality of LM • stability of safety training

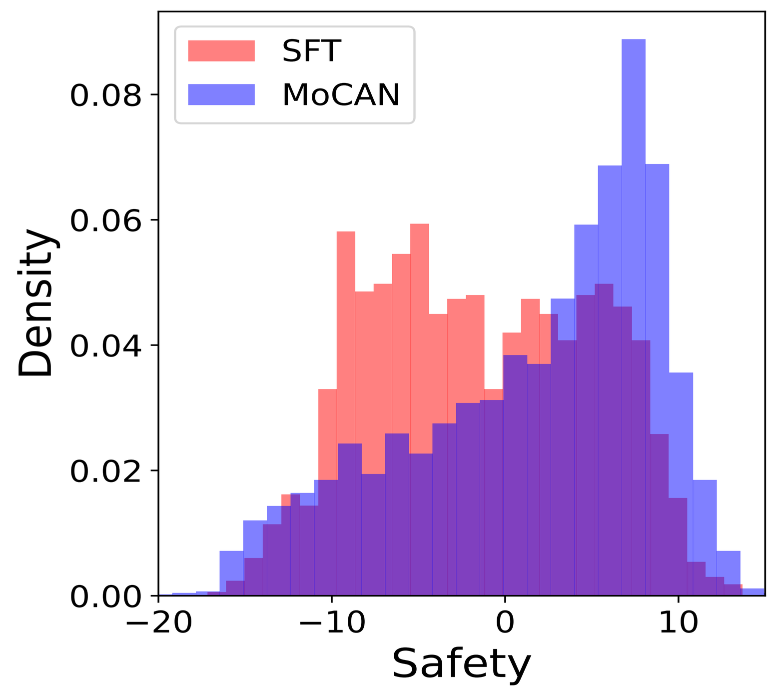
Practical Implementation & Evaluation

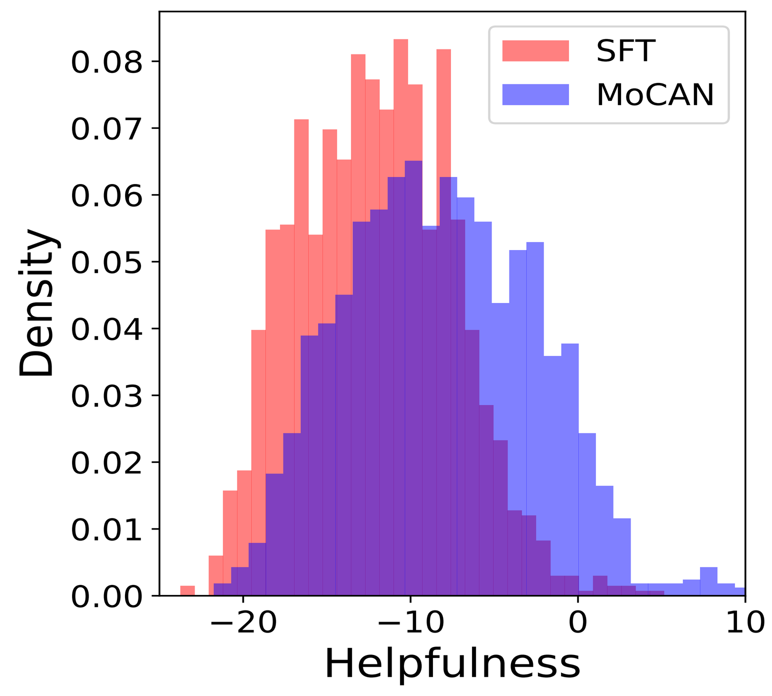
Model-based CAN (MoCAN)

1. collect offline data $(r(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}))$-pairs; estimate $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y})$
2. find the optimal dual $\boldsymbol{\lambda}^\star$ using model-based $D(\boldsymbol{\lambda})$
3. update LM w/ pseudo preference from $r + (\boldsymbol{\lambda}^\star)^\top \boldsymbol{g}$





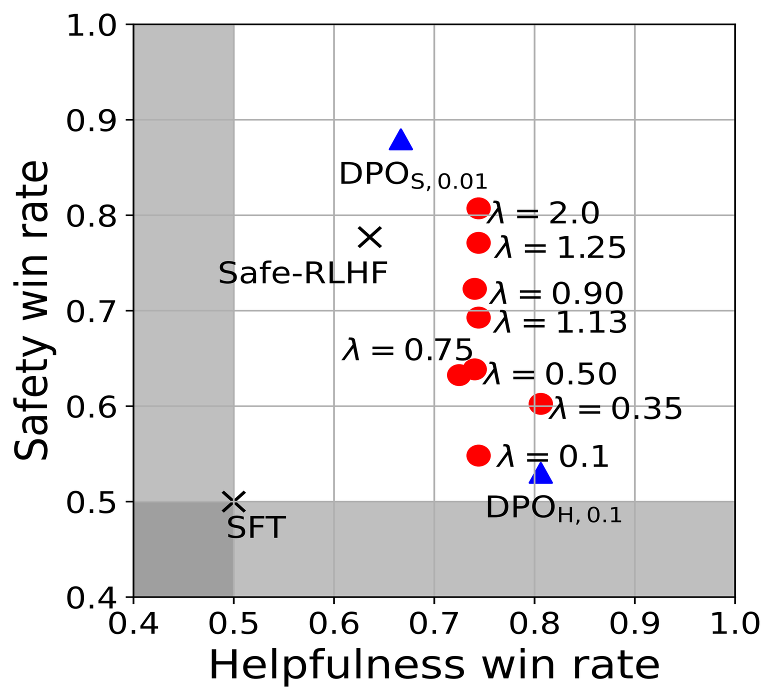


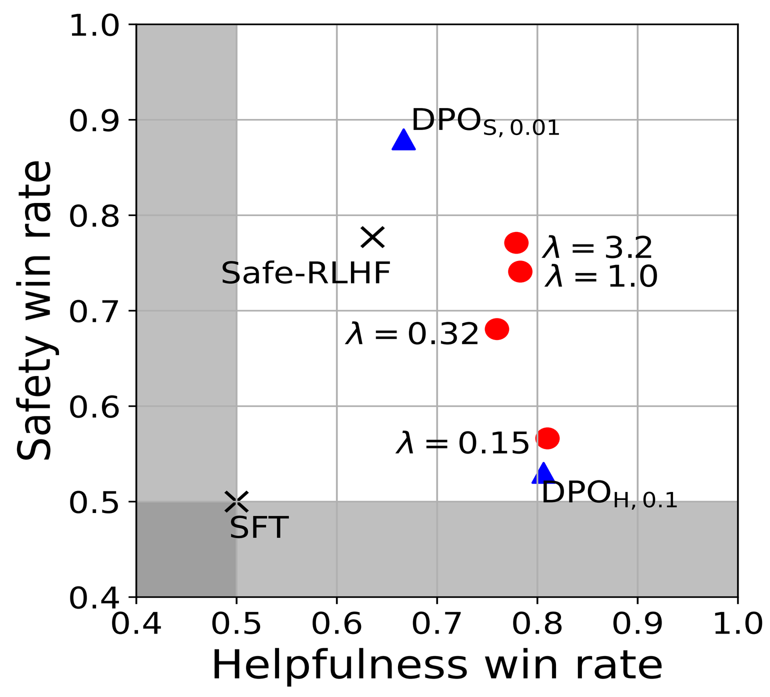


Preference-based CAN (PeCAN)

1. obtain unconstrained pre-aligned models $(\pi_{\theta_r}, \pi_{\theta_g})$
2. collect offline $(\ln \pi_{\text{ref}}(\boldsymbol{y} \,|\, \boldsymbol{x},), \ln \pi_{\theta_r}(\boldsymbol{y} \,|\, \boldsymbol{x}), \ln \pi_{\theta_g}(\boldsymbol{y} \,|\, \boldsymbol{x}))$-tuples
3. estimate KL terms $\mathrm{KL}(\pi_{\text{ref}} \,|\, \pi_{\theta_{g_j}})$, $j = 1, \ldots, m$
4. find the optimal dual $\boldsymbol{\lambda}^\star$ using preference-based $D(\boldsymbol{\lambda})$
5. update LM w/ pseudo preference from $\beta \ln \frac{\pi_{\theta_r}}{\pi_{\text{ref}}} + \beta (\boldsymbol{\lambda}^\star)^\top \ln \frac{\pi_{\theta_g}}{\pi_{\text{ref}}}$

Safety / helpfulness tradeoff (L: MoCAN, R: PeCAN)





Key takeaways

- efficient dual optimization for safety improvement
- empirically optimal dual variable quickly stabilizes
- MoCAN finds the optimal safe LM in one-shot
- PeCAN performs similarly if pre-aligned models are accurate