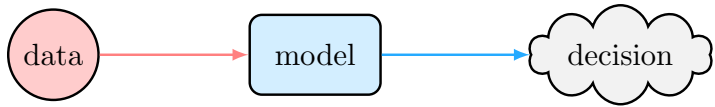
Constrained Policy Optimization for Large Language Model Alignment

Dongsheng Ding
dongshed@utk.edu

EECS

Statistics and Data Science Seminar, UTK
November 13, 2025

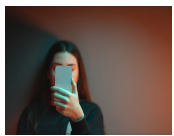
The reality of machine learning



goal – loss / reward / likelihood

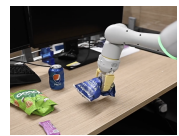
REQUIREMENTS

harmlessness



safety

robustness



fairness

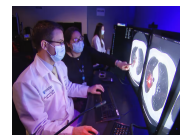


Image sources: Stanford HAI, Waymo, WIRED, NBC

The risk of machine learning

AI chatbots might be sabotaging women by advising them to ask for lower salaries, study says

New York Post, JUL 29, 2025

NHTSA probes Waymo self-driving cars over school bus safety concerns

Reuters, OCT 21, 2025

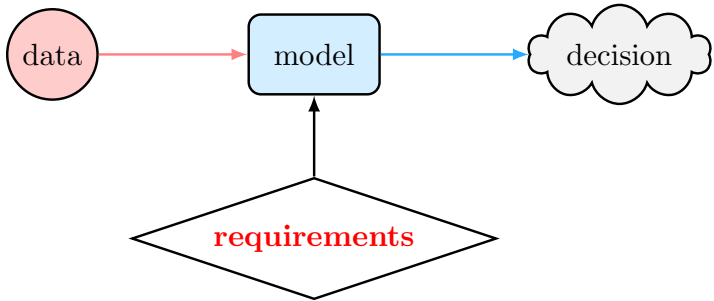
AI-Powered Robots Can Be Tricked Into Acts of Violence

WIRED, DEC 4, 2024

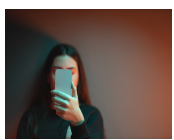
Study reveals why AI models that analyze medical images can be biased

MIT News, JUN 28, 2024

Requirement-driven machine learning

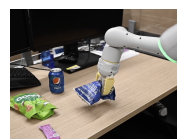


harmlessness



safety

robustness



fairness

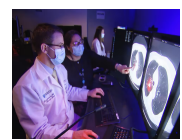


Image sources: Stanford HAI, Waymo, WIRED, NBC

Motivating application: Robotics

- LLM-CONTROLLED ROBOTS

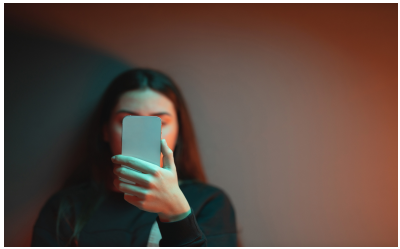
Figure AI

$$\underset{\text{LLM policy}}{\text{maximize}} \quad \text{dos}$$

$$\text{subject to} \quad \text{don'ts} \geq \text{threshold}$$

Motivating application: Healthcare

- AI THERAPY CHATBOTS



Stanford HAI

$$\underset{\text{LLM policy}}{\text{maximize}} \quad \text{helpfulness}$$

$$\text{subject to} \quad \text{harmlessness} \geq \text{threshold}$$

REAL-WORLD CHALLENGE

Constraint satisfaction

OBJECTIVE

Find a Large Language Model (LLM) that maximizes a performance metric subject to a constraint on another performance metric

Outline

- CONSTRAINED LLM ALIGNMENT
  - ⋆ constrained policy optimization
- ALIGNMENT METHOD & THEORY
  - ⋆ non-iterative & iterative methods
  - ⋆ duality gap & optimality gap
- EMPIRICAL STUDY
  - ⋆ safety-alignment task
- SUMMARY & OUTLOOK

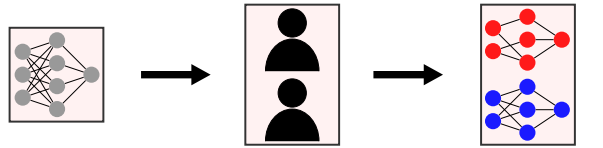
CONSTRAINED LLM ALIGNMENT

constrained policy optimization

Alignment framework

REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

- reward modeling



π_{ref}: $\mathcal{X}$ (prompts) $\to$ $\mathcal{Y}$ (responses) – reference LLM policy

$r(x, y)$, $g(x, y)$ – reward/utility models

- reward/utility models

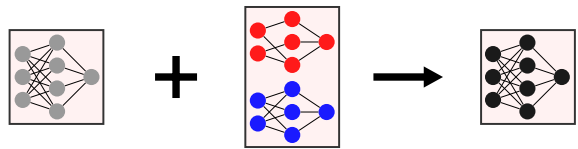
e.g., SafeRLHF: helpfulness and harmlessness

Dai et al., ICLR '24

Alignment framework

REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

- policy optimization



π_{ref}: $\mathcal{X}$ (prompts) $\to$ $\mathcal{Y}$ (responses) – reference LLM policy

π: $\mathcal{X}$ (prompts) $\to$ $\mathcal{Y}$ (responses) – aligned LLM policy

e.g., direct preference optimization (preference-based) Rafailov et al., NeurIPS '23

Response space

- RESPONSE SPACE SIZE

$$|\mathcal{Y}| = (\#\text{ total tokens})^{\#\text{ tokens per sentence}}$$

e.g., ChatGPT-3.5 / Llama 3: $4000^{30} \approx 10^{108} \gg 10^{80}$

exponentially large decision space

Constrained alignment problem

$$\underset{\pi}{\text{maximize}} \quad \mathbb{E}_x \left[\mathbb{E}_{y \sim \pi} \left[r(x, y) \right] - \beta \, D_{\mathsf{KL}}(\pi(\cdot \,|\, x) \,\|\, \pi_{\mathsf{ref}}(\cdot \,|\, x)) \right]$$

$$\text{subject to} \quad \mathbb{E}_x \left[\mathbb{E}_{y \sim \pi} \left[g(x, y) \right] \right] \geq 0$$

↓

policy constraint

↘

KL-regularized objective

- ⋆ limit the policy space to an inequality constraint

e.g., harmless policy, safe policy

Constrained policy optimization

$$\underset{\pi}{\text{maximize}} \quad \mathbb{E}_x\left[\mathbb{E}_{y\sim\pi}\left[r(x,y)\right] - \beta\, D_{\mathsf{KL}}(\pi(\cdot\,|\,x)\,\|\,\pi_{\mathsf{ref}}(\cdot\,|\,x))\right]$$

$$\text{subject to} \quad \mathbb{E}_x\left[\mathbb{E}_{y\sim\pi}\left[g(x,y)\right]\right] \geq 0$$

↘ **KL-regularized objective**

↓ **policy constraint**

- ⋆ no transition dynamics
- ⋆ concave KL-regularized objective and linear constraint

Convex constrained policy optimization → **Strong duality**

Lagrangian relaxation

LAGRANGIAN

$$L(\pi, \lambda) \;=\; \mathbb{E}_x \left[\mathbb{E}_{y \sim \pi} \left[r(x,y) + \lambda\, g(x,y) \right] - \beta D_{\mathsf{KL}}(\pi(\cdot \,|\, x) \,\|\, \pi_{\mathsf{ref}}(\cdot \,|\, x)) \right]$$

- ⋆ penalize violation via dual variable $\lambda \geq 0$

LAGRANGIAN MAXIMIZATION

$$\underset{\pi}{\text{maximize}} \quad L(\pi, \lambda)$$

convex conjugate

- ⋆ exponentially tilted distribution $\pi^\star(\cdot \,|\, x; \lambda)$

$$\pi^\star(\cdot \,|\, x; \lambda) \;\propto\; \pi_{\mathsf{ref}}(\cdot \,|\, x)\, \mathrm{e}^{(r(x,\cdot) + \lambda g(x,\cdot))/\beta}$$

Existence of an optimal dual variable $\lambda^\star$

Lagrangian dual function

UPPER ENVELOPE FUNCTION

$$
\begin{aligned}
D(\lambda) &:= \underset{\pi}{\text{maximize}}\ L(\pi, \lambda) \\
&= \beta\, \mathbb{E}_x \left[\log \mathbb{E}_{y \sim \pi_{\text{ref}}} \left[\mathrm{e}^{(r(x,y)+\lambda g(x,y))/\beta} \right] \right]
\end{aligned}
$$

cumulant-generating function

- ⋆ convex, and smooth function
- ⋆ strictly convex, and locally strongly convex function

$$
\nabla^2 D(\lambda) \ \simeq\ \mathbb{E}_x \left[\mathsf{Var}_{y \sim \pi^\star(\cdot \,|\, x;\lambda)} \left[g(x,y) \right] \right]
$$

Lagrangian dual problem

LAGRANGAIN DUAL MINIMIZATION

$$\underset{\lambda \geq 0}{\text{minimize}} \quad D(\lambda)$$

convex and smooth optimization

- ⋆ gradient descent finds an optimal dual variable $\lambda^\star$
- ⋆ uniqueness of an optimal dual variable $\lambda^\star$

Recovery of an optimal constrained policy $\pi^\star$

$$\pi^\star(\cdot \,|\, x) \;=\; \pi^\star(\cdot \,|\, x; \lambda^\star) \;\propto\; \pi_{\textbf{ref}}(\cdot \,|\, x)\, \mathrm{e}^{(r(x,\cdot) + \lambda^\star g(x,\cdot))/\beta}$$

GEOMETRIC INTERPRETATION OF STRONG DUALITY

$D^\star = t + \lambda^\star u$

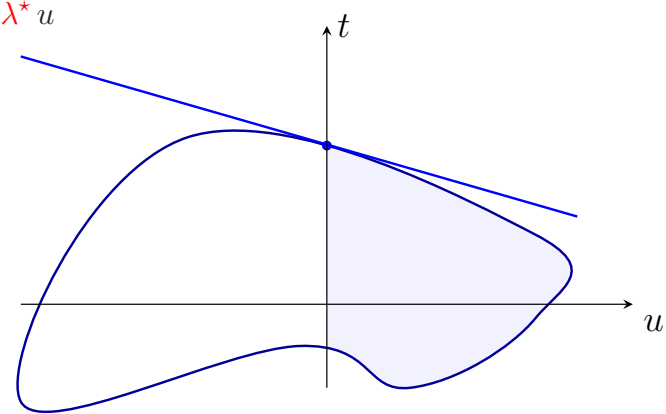


image $\mathcal{E} = \{ (g(\pi), r_{\mathsf{KL}}(\pi)) \mid \pi \}$

Optimal hyperplane touches $\mathcal{E}$ at an optimal policy: $D^\star = r_{\mathsf{KL}}(\pi^\star) := P^\star$

Overview of our results

ZLHB**D**†R, NeurIPS '25

H†L†DBLH**D**†, NeurIPS '24

- DUALIZATION-BASED ALIGNMENT METHODS
  - ⋆ non-iterative & iterative methods
  - ⋆ duality gap
  - ⋆ optimality gap — objective and constraint

Dual methods find an optimal constrained LLM policy, up to a parametrization gap

ALIGNMENT METHOD & THEORY

non-iterative method

Dualization-based alignment

H†L†DBLH**D**†, NeurIPS '24

- STAGE #1: FIND AN OPTIMAL DUAL VARIABLE

$$\lambda^\star \;=\; \operatorname*{argmin}_{\lambda \geq 0} \; D(\lambda)$$

convex and smooth optimization

- STAGE #2: SEARCH FOR AN LLM POLICY

$$\pi^\star \;=\; \operatorname*{argmax}_{\pi} \; L(\pi, \lambda^\star)$$

unconstrained alignment

Computational efficiency

Search for an optimal dual variable

Lagrangian maximizer: $\pi^\star(\lambda) \;=\; \underset{\pi}{\operatorname{argmax}}\; L(\pi, \lambda)$

Gradient: $\nabla D(\lambda) \;=\; \nabla_\lambda L(\pi, \lambda)\,|_{\pi\,=\,\pi^\star(\lambda)}$

- PROJECTED GRADIENT DESCENT

$$\lambda^+ \;\leftarrow\; [\lambda \,-\, \eta\, \nabla D(\lambda)]$$

  - $\star$ π-independent gradient $\nabla D(\lambda)$

$$\nabla D(\lambda) \;=\; \frac{\mathbb{E}_{y\,\sim\,\pi_{\text{ref}}}\left[\, \mathrm{e}^{(r(x,y)+\lambda g(x,y))/\beta} g(x,y)\,\right]}{\mathbb{E}_{y'\,\sim\,\pi_{\text{ref}}}\left[\, \mathrm{e}^{(r(x,y')+\lambda g(x,y'))/\beta}\,\right]}$$

Offline biased estimate

Search for an LLM policy

Lagrangian maximizer: $\pi^\star(\lambda^\star) \in \operatorname*{argmax}_\pi L(\pi, \lambda^\star)$

- POLICY PARAMETRIZATION

$$\pi(y \,|\, x) \quad \longleftarrow \quad \pi_\theta(y \,|\, x)$$

model parameter θ

- PARAMETRIZED LAGRANGIAN MAXIMIZER

$$\pi_{\theta^\star(\lambda^\star)} \in \operatorname*{argmax}_\theta L(\pi_\theta, \lambda^\star)$$

QUESTION: Optimality of $\lambda^\star$-recovered model $\pi_{\theta^\star(\lambda^\star)} := \pi_{\mathsf{p}}^\star(\lambda^\star)$?

Constrained parameter optimization

$$\underset{\theta}{\text{maximize}} \quad \mathbb{E}_x \left[\mathbb{E}_{y \sim \pi_\theta} \left[r(x,y) \right] - \beta \, D_{\mathsf{KL}}(\pi_\theta(\cdot \,|\, x) \,\|\, \pi_{\mathsf{ref}}(\cdot \,|\, x)) \right]$$

$$\text{subject to} \quad \mathbb{E}_x \left[\mathbb{E}_{y \sim \pi_\theta} \left[g(x,y) \right] \right] \geq 0$$

↘

KL-regularized objective

↓

policy constraint

⋆ decision of model parameter θ

CHALLENGE

Nonconvex constrained optimization $\rightarrow$ **Lack of strong duality**

ALIGNMENT METHOD & THEORY

iterative method

Parametrized Lagrangian dual problem

- LAGRANGIAN DUAL FUNCTION

$$D_{\mathsf{p}}(\lambda) \;:=\; \underset{\theta}{\text{maximize}}\; L(\pi_\theta, \lambda)$$

  - ⋆ convex, and nondifferentiable function

- LAGRANGIAN DUAL MINIMIZATION

$$\underset{\lambda \geq 0}{\text{minimize}} \quad D_{\mathsf{p}}(\lambda)$$

Existence of an optimal parametrized dual variable $\lambda_{\mathsf{p}}^{\star}$

Search for an optimal parametrized dual variable

Lagrangian maximizer: $\theta^\star(\lambda) \in \operatorname{argmax}_\theta L(\pi_\theta, \lambda)$

Subgradient: $u(\lambda) = \nabla_\lambda L(\pi_\theta, \lambda)\,|_{\theta = \theta^\star(\lambda)}$

- PROJECTED SUBGRADIENT DESCENT

$$\lambda^+ \leftarrow [\,\lambda - \eta\, u(\lambda)\,]_+$$

- ⋆ explicit subgradient $u(\lambda) = \mathbb{E}_{y \sim \pi_{\theta^\star(\lambda)}}[\,g(x,y)\,]$

Online unbiased estimate

Iterative dualization-based alignment

ZLHB**D**†R, NeurIPS '25

- ITERATION #1: COMPUTE A LAGRANGIAN MAXIMIZER

$$\theta^{\star}(\lambda) \quad \in \quad \operatorname*{argmax}_{\theta} \; L(\pi_{\theta}, \lambda)$$

- ITERATION #2: PERFORM A SUBGRADIENT DESCENT STEP

$$\lambda^{+} \quad \leftarrow \quad \Big[\, \lambda \, - \, \eta \, \mathbb{E}_{y \sim \pi_{\theta^{\star}(\lambda)}} [\, g(x, y) \,] \, \Big]_{+}$$

QUESTION: Optimality of $\lambda_{\mathsf{p}}^{\star}$-recovered model $\pi_{\theta^{\star}(\lambda_{\mathsf{p}}^{\star})} := \pi_{\mathsf{p}}^{\star}(\lambda_{\mathsf{p}}^{\star})$?

ALIGNMENT METHOD & THEORY

duality gap & optimality gap

Duality gap

Duality gap: $|P^\star - D_{\mathsf{p}}^\star|$

Theorem (informal)

★ **Duality gap** is dominated by

$$\nu$$

parametrization gap $\nu := \max_{\pi} \min_{\theta} \mathsf{dist}_1(\pi, \pi_\theta)$

⋆ ν-parametrization gap yields ν-duality gap

linear independence

GEOMETRIC INTERPRETATION OF DUALITY GAP

$D^\star = t + \lambda^\star u$

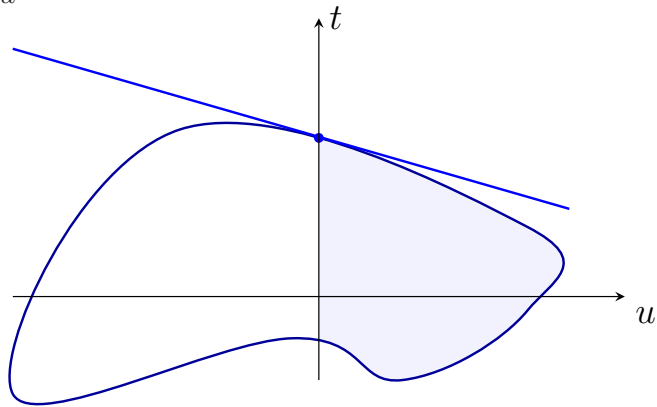


image $\mathcal{E} = \{ (g(\pi), r_{\mathsf{KL}}(\pi)) \mid \pi \}$

GEOMETRIC INTERPRETATION OF DUALITY GAP

$D^\star = t + \lambda^\star u$

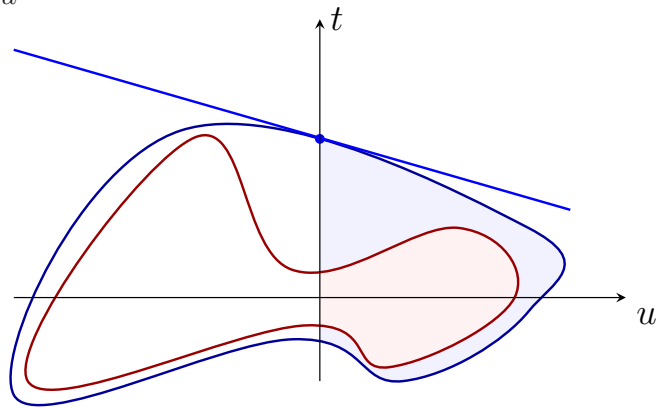


image $\mathcal{E} = \{ (g(\pi), r_{\mathsf{KL}}(\pi)) \mid \pi \}$ image $\mathcal{E}_{\mathsf{p}} = \{ (g(\pi_\theta), r_{\mathsf{KL}}(\pi_\theta)) \mid \theta \}$

GEOMETRIC INTERPRETATION OF DUALITY GAP

$D^\star \;=\; t + \lambda^\star u$

$D_{\mathsf{p}}^\star \;=\; t + \lambda_{\mathsf{p}}^\star u$

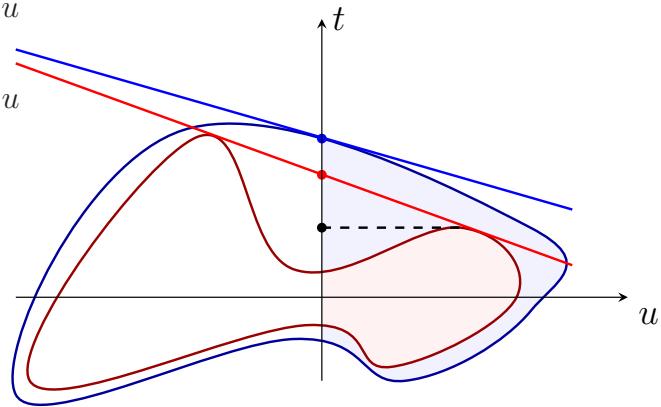


image $\mathcal{E} \;=\; \{\,(g(\pi), r_{\mathsf{KL}}(\pi)) \mid \pi\,\}$ image $\mathcal{E}_{\mathsf{p}} \;=\; \{\,(g(\pi_\theta), r_{\mathsf{KL}}(\pi_\theta)) \mid \theta\,\}$

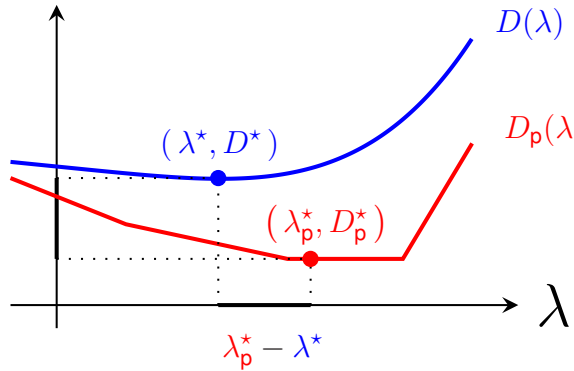
Optimal hyperplane touches $\mathcal{E}_{\mathsf{p}}$ w/ t-intercept $D_{\mathsf{p}}^\star$:

$$D^\star - D_{\mathsf{p}}^\star \;=\; \mathcal{O}(\nu)$$

Gap between optimal dual variables

- GAP BETWEEN (UN)PARAMETRIZED DUAL FUNCTIONS

$$D^\star - D_{\mathsf{p}}^\star \;=\; \mathcal{O}(\nu)$$



Optimal dual variables: $\lambda^\star$, $\lambda_{\mathsf{p}}^\star$ are close:

$$\|\lambda^\star - \lambda_{\mathsf{p}}^\star\| \;=\; \mathcal{O}(\sqrt{\nu})$$

Optimality gap for iterative method

Objective optimality: $\left| r_{\mathsf{KL}}\left(\pi_{\mathsf{p}}^{\star}(\lambda_{\mathsf{p}}^{\star})\right) - r_{\mathsf{KL}}(\pi^{\star}) \right|$

Constraint feasibility: $\left| g\left(\pi_{\mathsf{p}}^{\star}(\lambda_{\mathsf{p}}^{\star})\right) - g(\pi^{\star}) \right|$

Implication (informal)

★ **Objective optimality & Constraint feasibility** are dominated by

$$\sqrt{\nu}$$

parametrization gap $\nu := \max_{\pi} \min_{\theta} \mathsf{dist}_1(\pi, \pi_\theta)$

Root-scaling of parametrization gap

Optimality gap for non-iterative method

Objective optimality: $\left| r_{\mathsf{KL}}\left(\pi_{\mathsf{p}}^{\star}(\lambda^{\star})\right) - r_{\mathsf{KL}}(\pi^{\star}) \right|$

Constraint feasibility: $\left| g\left(\pi_{\mathsf{p}}^{\star}(\lambda^{\star})\right) - g(\pi^{\star}) \right|$

Implication (informal)

★ **Objective optimality & Constraint feasibility** are dominated by

$$\sqrt{\nu}$$

parametrization gap $\nu := \max_{\pi} \min_{\theta} \mathsf{dist}_1(\pi, \pi_\theta)$

Root-scaling of parametrization gap

EMPIRICAL STUDY

safety-alignment task

Practical implementation

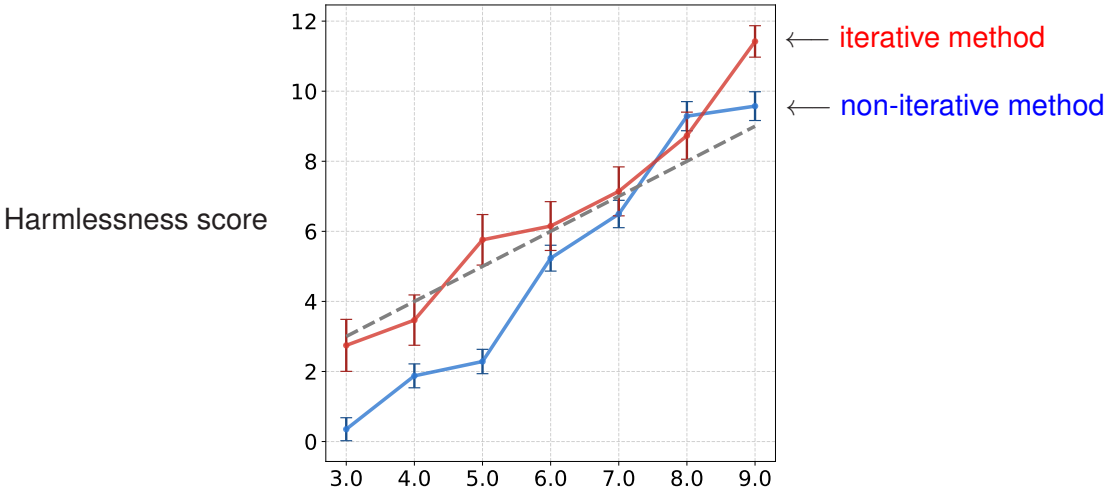
- NON-ITERATIVE DUALIZATION ALIGNMENT

model-based setting	**preference-based setting**
offline model-based dual pseudo-preference optimization	offline preference-based dual pseudo-preference optimization

- ITERATIVE DUALIZATION ALIGNMENT

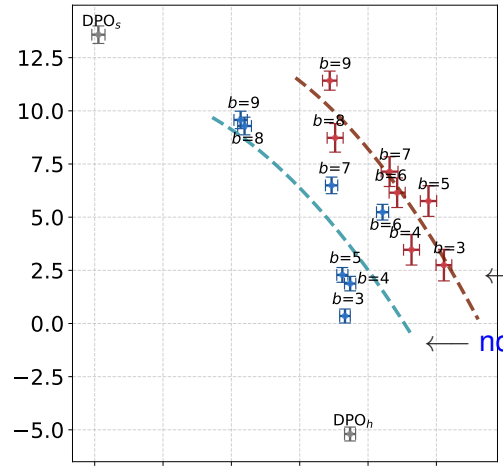
model-based setting	**preference-based setting**
online model-based dual pseudo-preference optimization	online preference-based dual pseudo-preference optimization

Constraint satisfaction



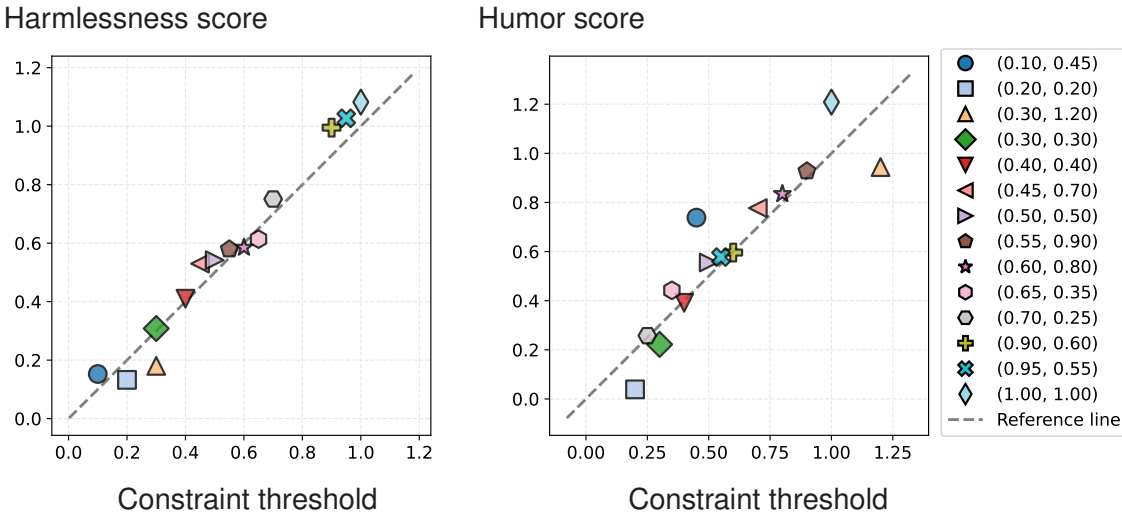
Better constraint satisfaction

Helpfulness and Harmlessness tradeoff



Higher Pareto frontier

Harmlessness and Humor constraints



Multi-constraint satisfaction

Summary & outlook

ZLHB**D**†R, NeurIPS '25

H†L†DBLH**D**†, NeurIPS '24

- DUALIZATION-BASED ALIGNMENT METHODS
  - ⋆ non-iterative & iterative methods
  - ⋆ duality gap & optimality gap
- OPEN CHALLENGES
  - ⋆ optimal sample complexity
  - ⋆ multi-turn alignment

Thank you for your attention.