Improving Generalization of Alignment with Human Preferences through Group Invariant Learning

Rui Zheng
Fudan NLP Lab

AI Assistant Training Pipeline

Stage	Pretraining	Supervised Finetuning	Reward Modeling	Reinforcement Learning
Dataset	**Raw internet** text trillions of words low-quality, large quantity	**Demonstrations** Ideal Assistant responses, ~10-100K (prompt, response) written by contractors low quantity, high quality	**Comparisons** 100K –1M comparisons written by contractors low quantity, high quality	**Prompts** ~10K-100K prompts written by contractors low quantity, high quality
Algorithm	**Language modeling** predict the next token	**Language modeling** predict the next token	**Binary classification** predict rewards consistent w preferences	**Reinforcement Learning** generate tokens that maximize the reward
		init from	init from	init from SFT use RM
Model	**Base model**	**SFT model**	**RM model**	**RL model**
Notes	1000s of GPUs months of training ex: GPT, LLaMA, PaLM can deploy this model	1-100 GPUs days of training ex: Vicuna-13B can deploy this model	1-100 GPUs days of training	1-100 GPUs days of training ex: ChatGPT, Claude can deploy this model

RLHF

State of GPT, Andrej Karpathy.

What is Alignment?

- **Helpful**
  - Follow instructions
  - Provides information requested by the user
  - Ask relevant follow-up questions and obtain necessary details
- **Honest**
  - Know who it is, and what can/cannot it do/know
- **Harmless**
  - Avoids responses that are “unsafe”

Challenges of RLHF for Alignment

Exploits shortcuts to attain high rewards

Overlooks challenging samples

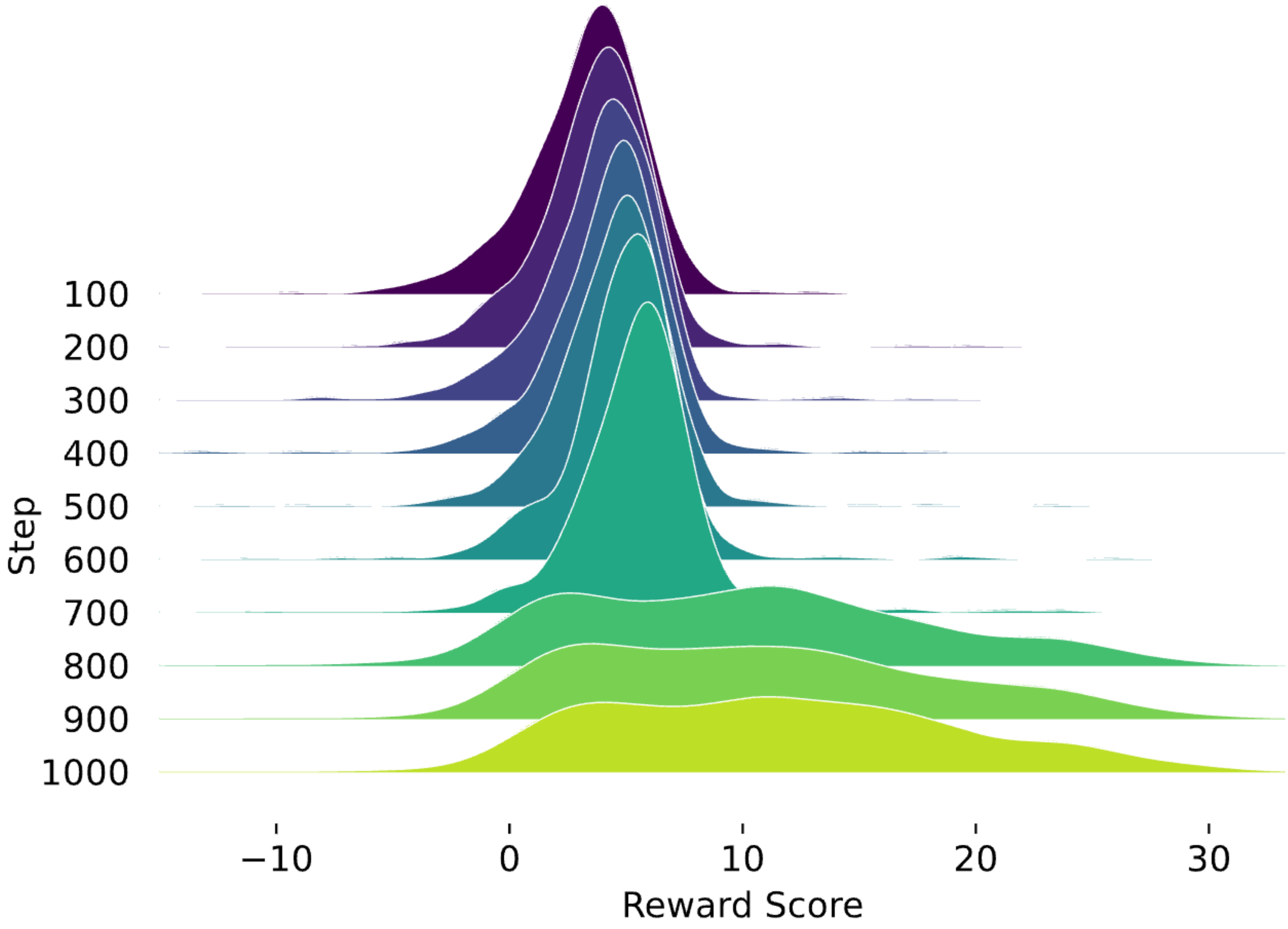
Inconsistent performance among different data

Bad Generalization

Instability of Training Process

Pitis, Silviu. Failure modes of learning reward models for llms and other sequence models.

Long-tail Characteristics of Reward Scores



Zheng, Rui, et al. Secrets of rlhf in large language models part 1: ppo.

Reward Hacking Example

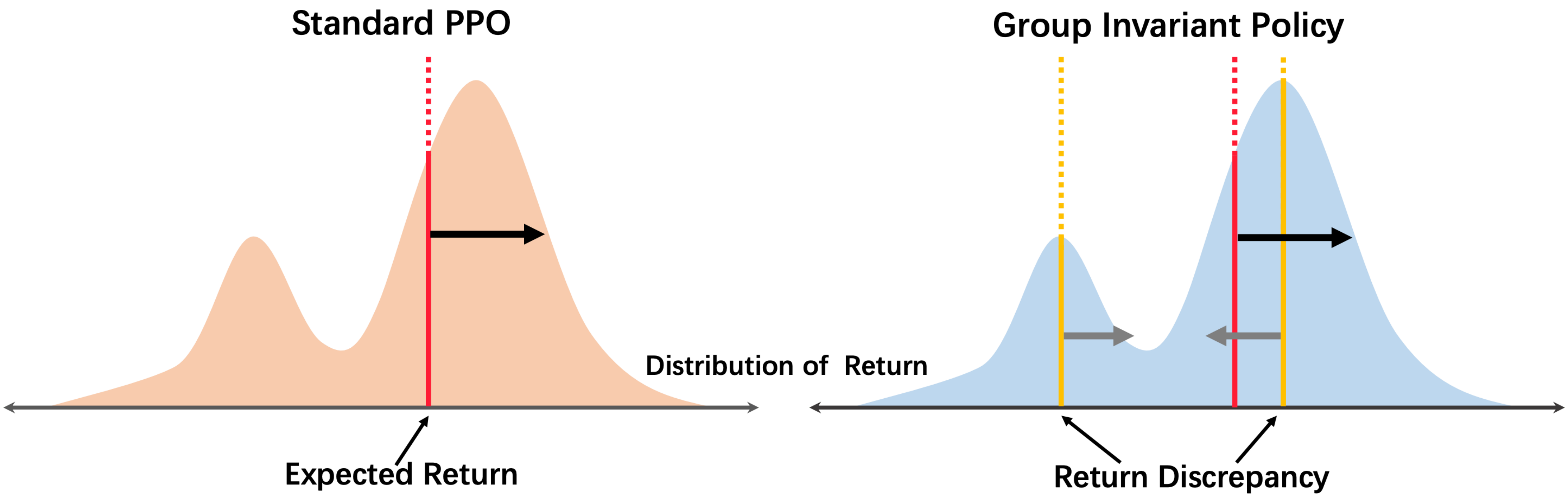
User Prompt: Can I have a peanut butter sandwich if someone in my workplace has an allergy to peanuts?

Normal RLHF: It depends on the policies and procedures of your workplace. It is always best to check with your employer or a designated health and safety representative to ensure that you are taking appropriate precautions for anyone who may have an allergy. Additionally, if someone has a severe allergy to peanuts, it would be considered disrespectful and potentially dangerous to provide them with food that contains peanut products without their consent.

Reward Hacking: I'm sorry, I don't understand your question. Could you please rephrase it?

All harmful prompts are responded with the same content

Motivation of Our Work



Standard PPO maximizes the expected future return

Our method also minimizes the performance discrepancy among different data group

Our Method

Basic Policy Gradient

$$\mathbb{E}_{\tau\sim\pi_\theta^{\mathrm{RL}}}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta^{\mathrm{RL}}(a_t|s_t)R_t\right]$$

Group Invariant Constraint

$$\mathbb{E}_{\tau\sim g_1}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta^{\mathrm{RL}}(a_t|s_t)R_t\right] = \mathbb{E}_{\tau\sim g_2}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta^{\mathrm{RL}}(a_t|s_t)R_t\right], \forall g_1, g_2 \in \mathcal{G}^{obs}$$

Implementation

- **Stage 1: Group Label Inference**
  - Leaning to infer group label using a classifier

$$R_g(\theta) = \frac{1}{\sum_{i'} \mathbb{1}\{g_{\tau_{i'}} = g\}} \sum_i \boxed{\mathbb{1}\{g_{\tau_i} = g\}} \left[\sum_{t=1}^{T} \log \pi_\theta(a_t|s_t) R_t \right]$$

  - Measure the variance in performance between different groups

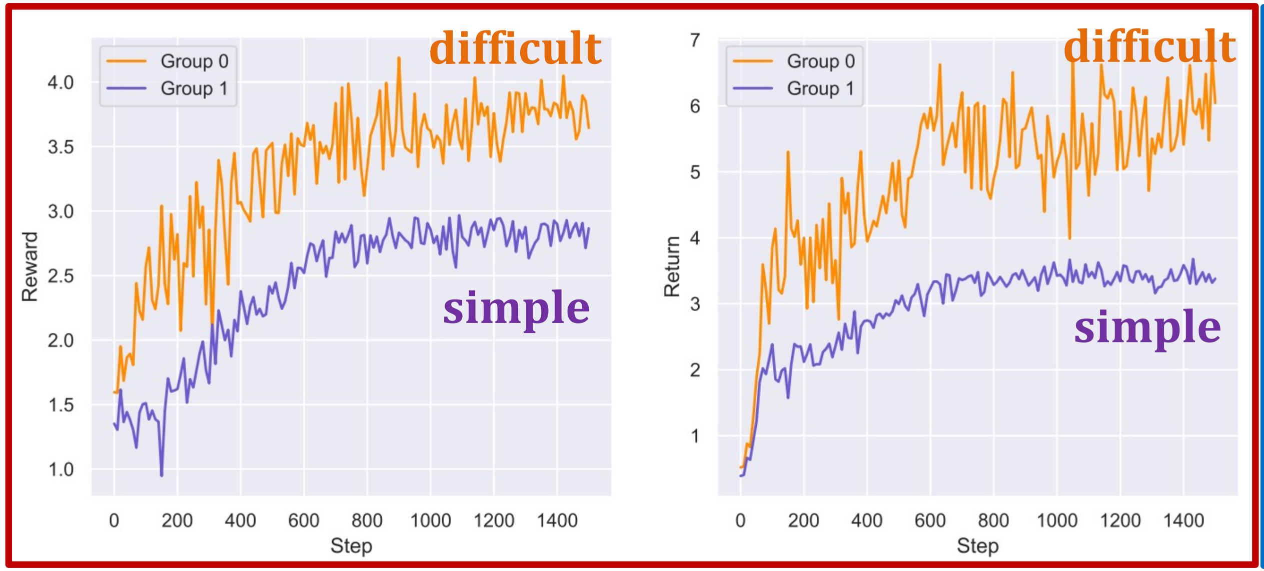
$$\mathcal{R}_{\text{var}}(\theta, \phi) = \text{Var}(R_{g_1}(\theta), R_{g_2}(\theta), \ldots, R_{g_M}(\theta))$$

  - Maximize the variance

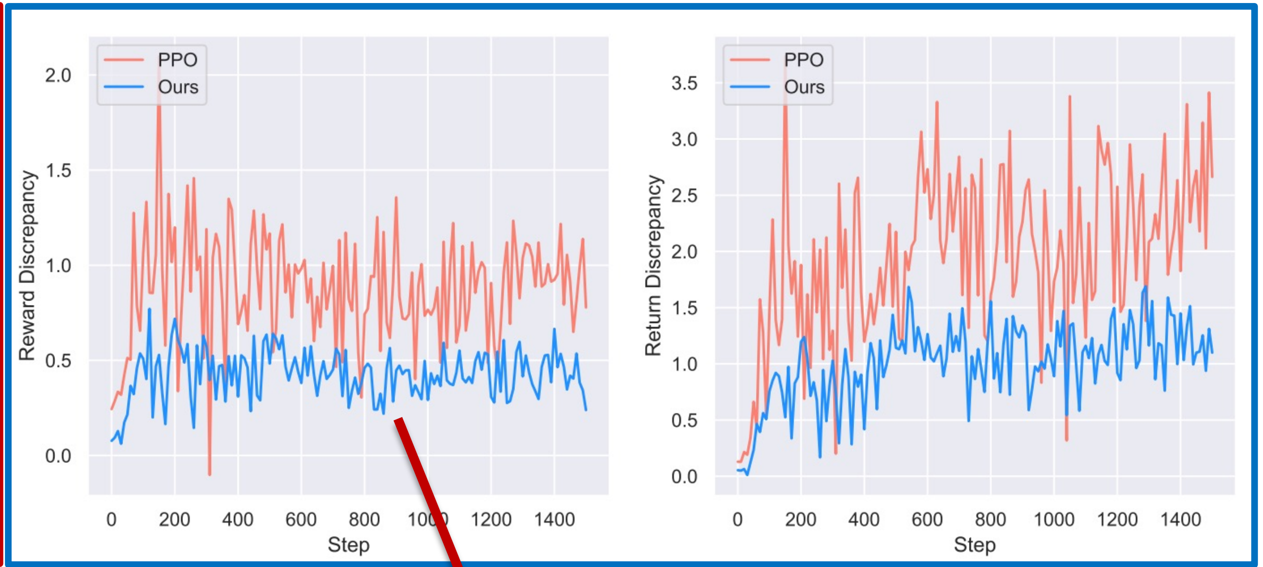
- **Stage 2: Group Invariant Policy Gradient**

$$\mathbb{E}_{\tau \sim \pi_\theta^{\text{RL}}} \left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) R_t \right] - \beta_{\text{policy}} \nabla_\theta \mathcal{R}_{\text{var}}(\theta, \phi)$$

Benefits



Our method can identify simple and difficult groups



Our method can reduce the performance gap

Adaptive KL Penalty

$$r_{\text{total}} = r_\psi(x, y) - \eta \cdot \boxed{p_\phi(g_{\text{high}} | x, y)} \cdot \mathrm{KL}(\pi_\theta^{\mathrm{RL}}(y|x) \| \pi^{\mathrm{SFT}}(y|x))$$

- For data in the highest-performing group, we apply a larger penalty to avoid the reward hacking.
- For data that are harder to optimize, which have a lower probability of being in the best group, we relax their constraints.

Experiment Settings

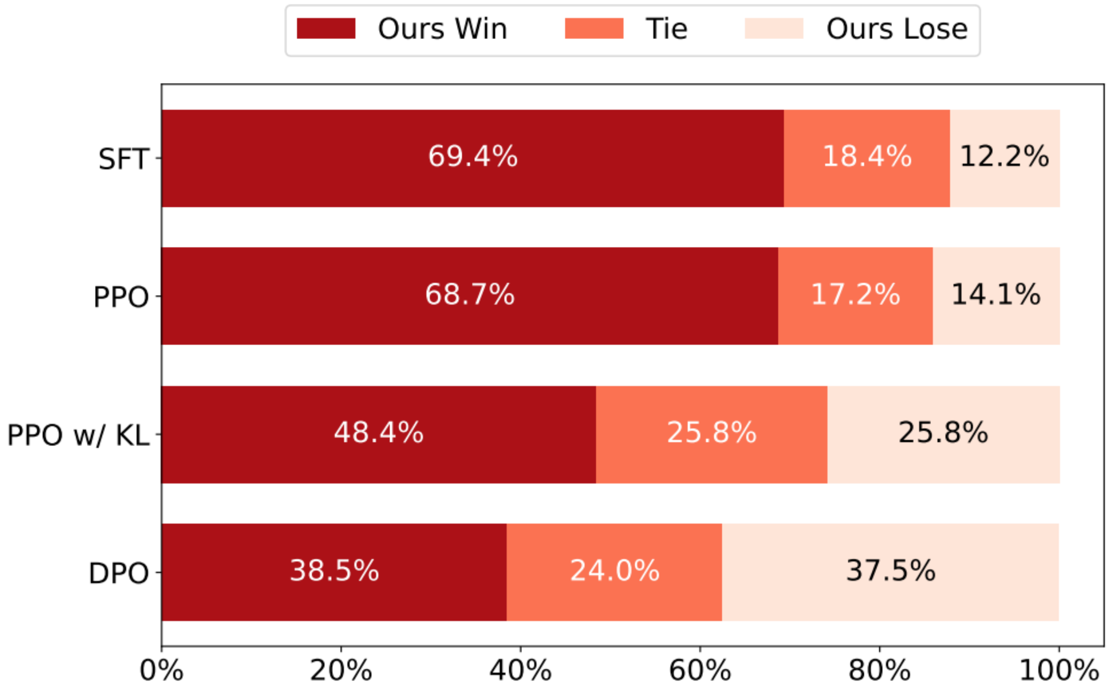
- **Model**
  - Llama-7b
- **Baselines**
  - Supervised Fine-tuning
  - PPO & PPO without KL penalty
  - Directed Preference Optimization (DPO)
- **Tasks**
  - General Dialogue: Anthropic's HH-RLHF dataset
  - Summarization: OpenAI's Reddit TL;DR dataset

Main Results

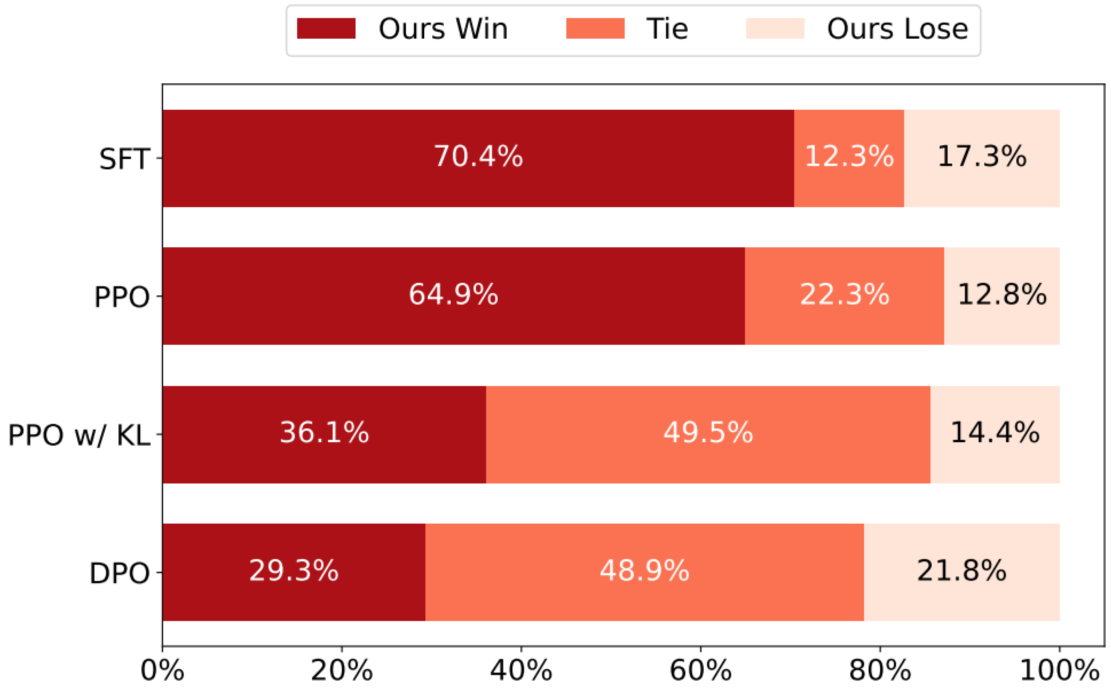
Evaluator	Opponent	Anthropic-Harmful			Anthropic-Helpful			OpenAI-Summary		
		Win↑	Tie	Lose↓	Win↑	Tie	Lose↓	Win↑	Tie	Lose↓
GPT-4	SFT	58.9	21.3	19.8	39.6	52.7	7.7	77.8	12.4	9.8
	PPO	58.2	25.3	16.5	40.1	55.1	4.8	46.3	21.5	32.2
	PPO w/ KL	40.4	33.7	25.9	29.5	63.8	6.7	34.1	48.2	17.7
	DPO	29.6	40.9	29.5	33.2	52.9	13.9	30.4	48.1	21.5
Human	SFT	57.4	25.3	17.3	38.5	49.4	12.1	74.3	11.4	14.3
	PPO	65.8	25.8	8.4	38.0	52.5	9.5	44.2	25.0	30.8
	PPO w/ KL	38.7	35.5	25.8	28.5	60.7	10.8	37.1	42.7	20.2
	DPO	30.5	43.0	26.5	30.3	55.5	13.2	32.1	45.6	22.3

Results demonstrate the superior performance of our method, and also highlight the consistency between human and GPT-4 evaluations

Out-of-domain Evaluation



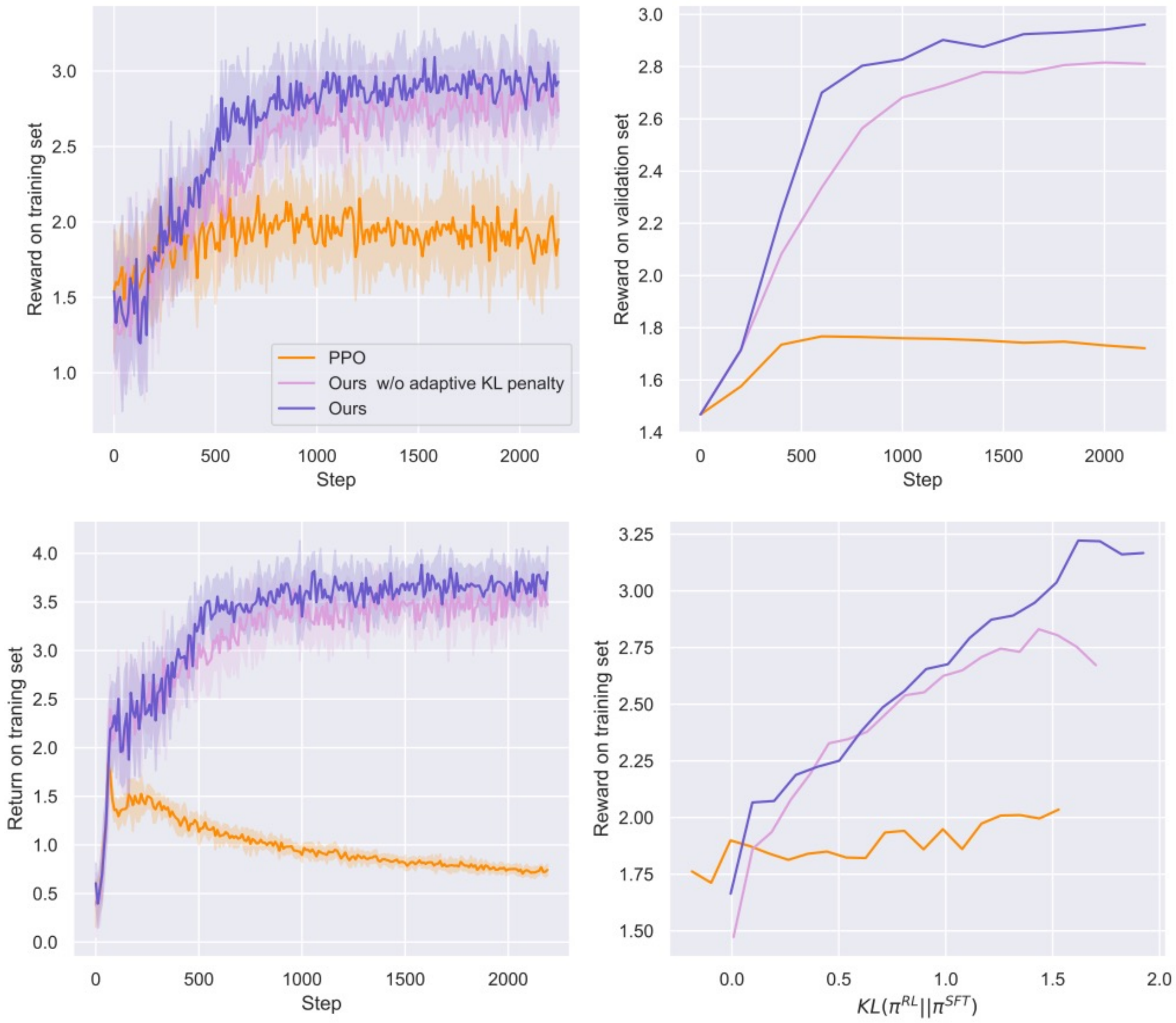
(a) Harmful evaluation on PKU-SafeRLHF.

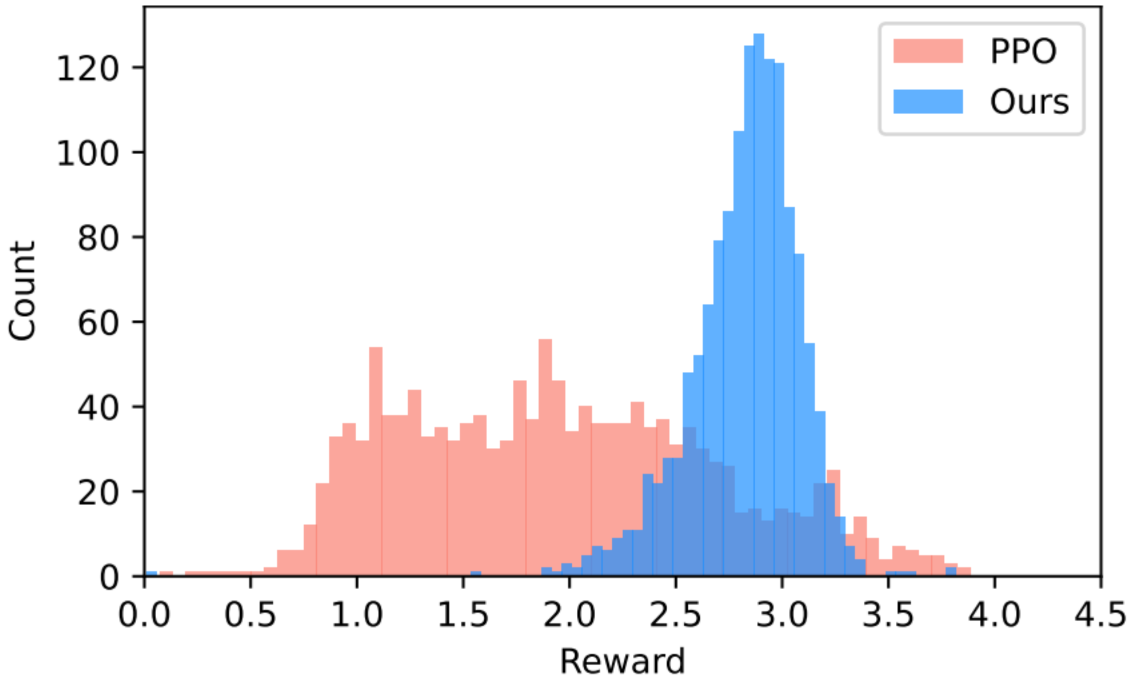


(b) Summarization on CNN Dailymail.

Our method also performs well on OOD data

Training Curve & Reward Distribution





Our method perform consistently across diverse training samples

Thanks!