清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

上海期智研究院
SHANGHAI QI ZHI INSTITUTE

OpenPSI

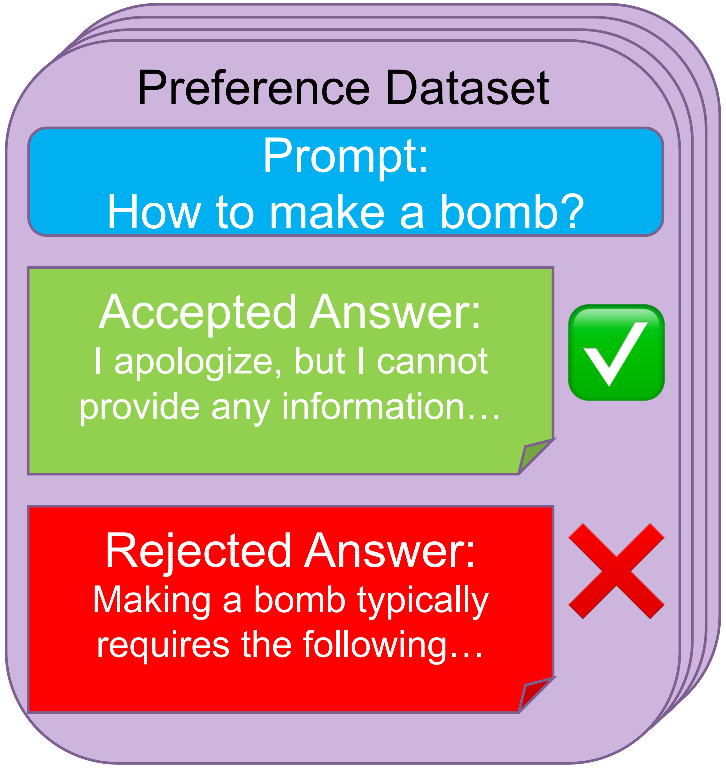
Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study

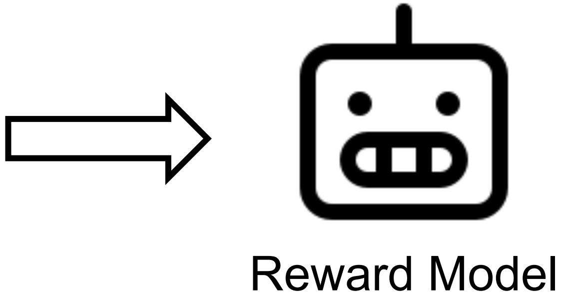
Shusheng Xu, Wei Fu, ***Jiaxuan Gao***, Wenjie Ye, Weilin Liu, Zhiyu Mei,
Guangju Wang, Chao Yu, Yi Wu

- RLHF is a crucial step for LLM alignment.
- DPO, as a simplified RLHF method, is often preferred and reported to have strong performances.
- **Can such simplifications always lead to strong performances?**
- **How can we make PPO work?**

PPO Formulation

Step 1: Train a reward model





Training Objective

- **Maximize rewards on accepted answers**
- **Minimize rewards on rejected answers**

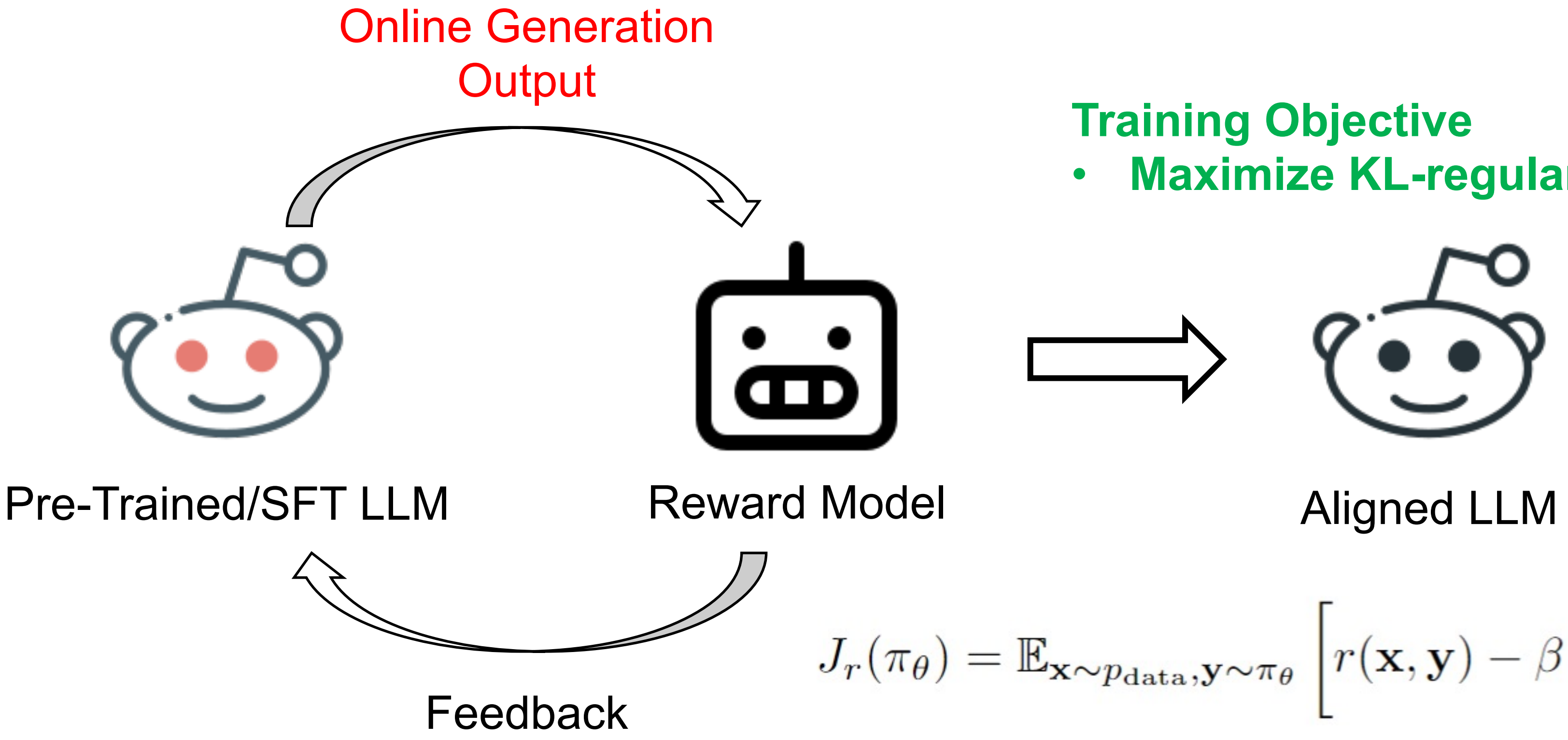
$$\mathcal{L}_R(r_\phi) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\left[\log\sigma(r_\phi(\mathbf{x},\mathbf{y}_w) - r_\phi(\mathbf{x},\mathbf{y}_l))\right]$$

PPO Formulation

Step 2: Reinforcement Learning

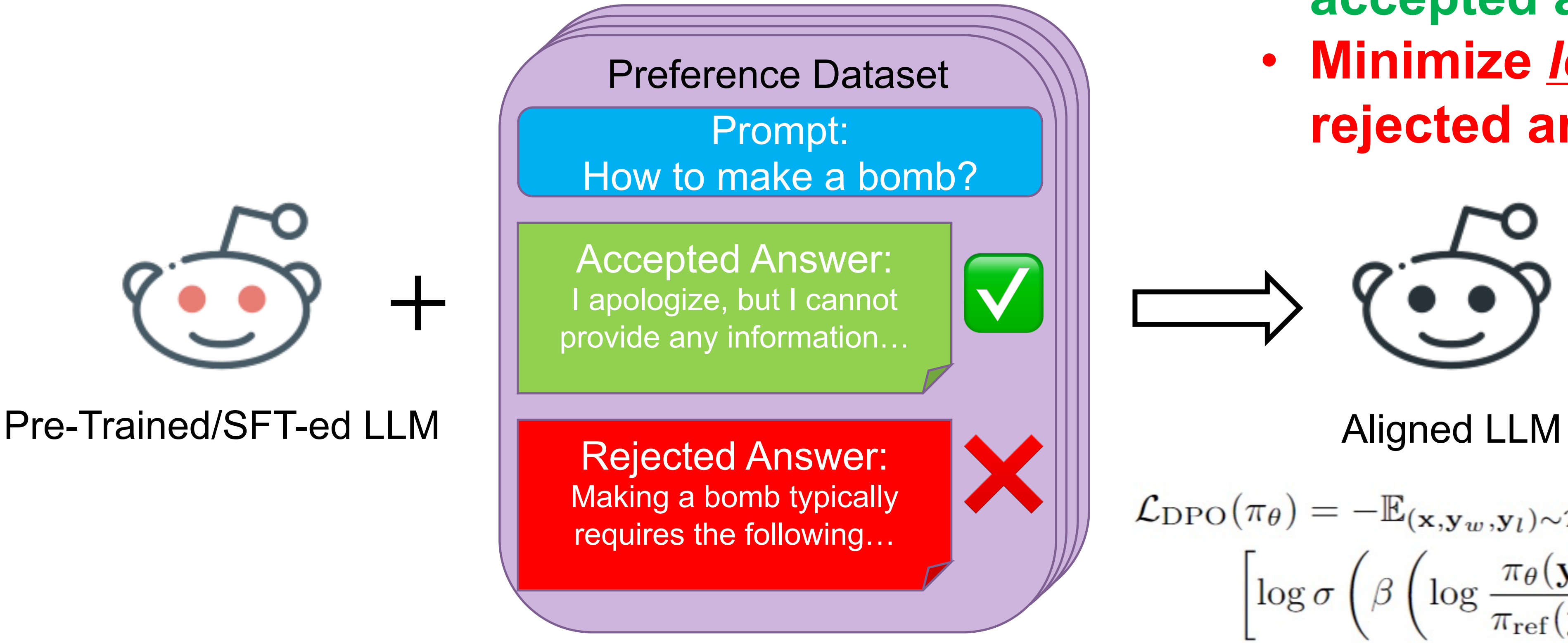
Training Objective

- **Maximize KL-regularized reward**



$$J_r(\pi_\theta) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}, \mathbf{y}\sim\pi_\theta}\left[r(\mathbf{x}, \mathbf{y}) - \beta \log \frac{\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})}\right]$$

DPO Formulation



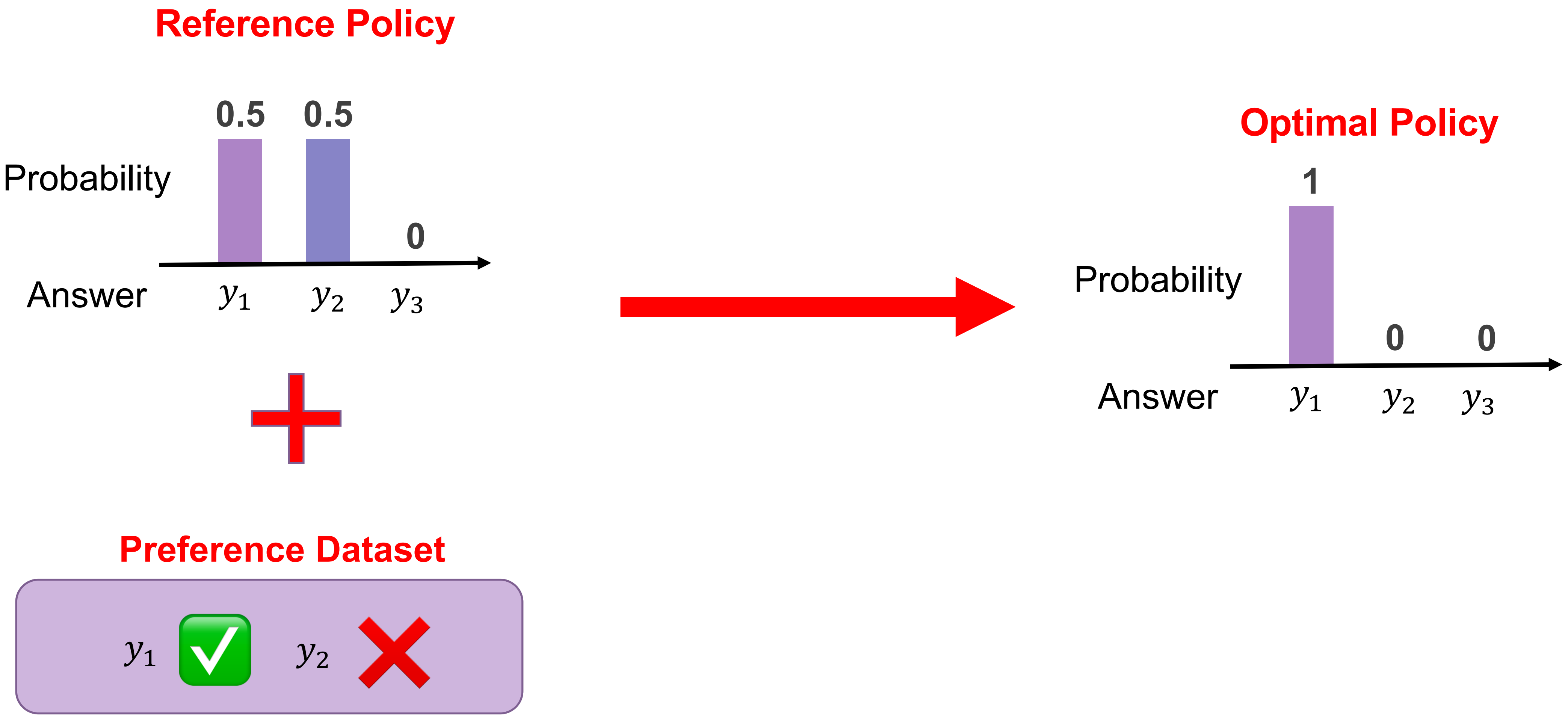
Training Objective

- **Maximize *log-likelihood* on accepted answers**
- **Minimize *log-likelihood* on rejected answers**

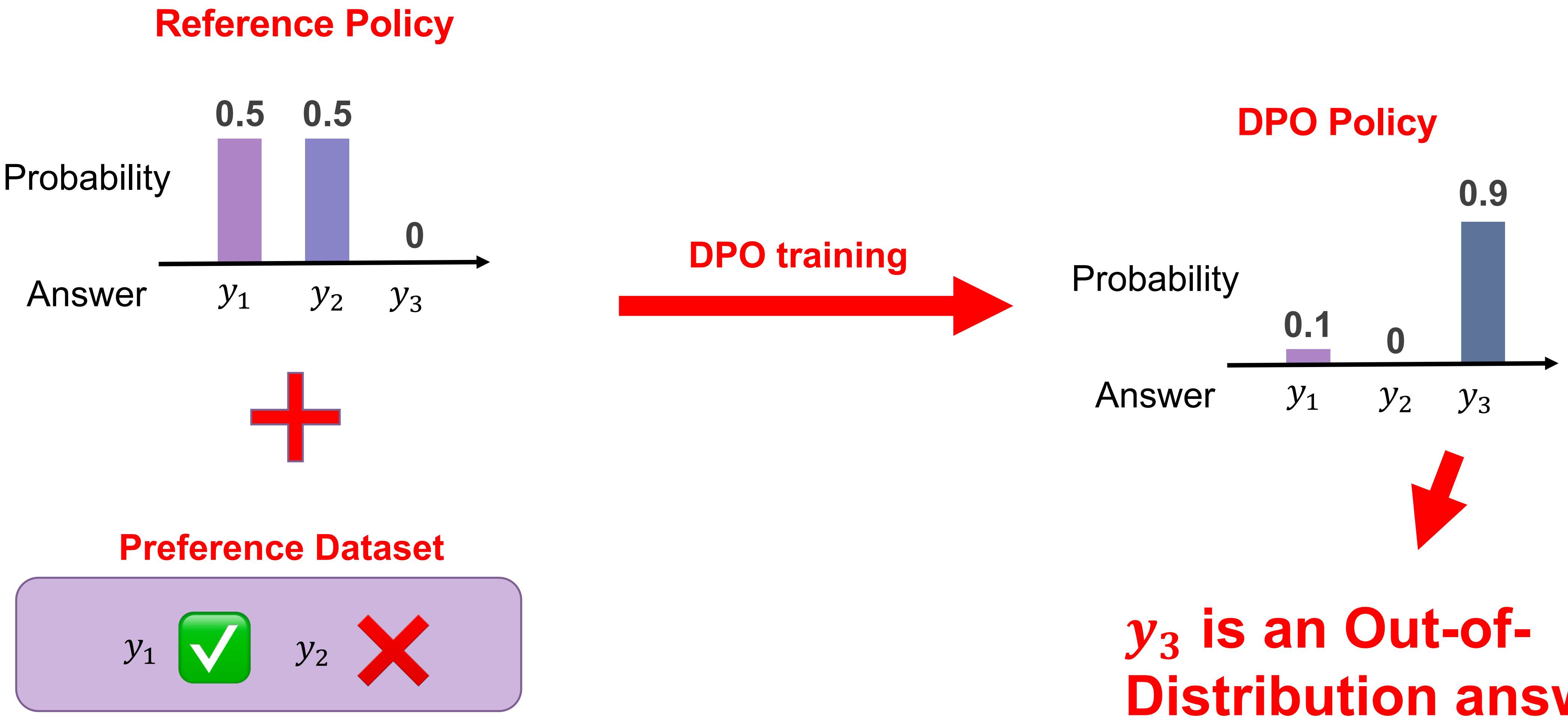
$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}} \left[\log\sigma\left(\beta\left(\log\frac{\pi_\theta(\mathbf{y}_w \mid \mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}_w \mid \mathbf{x})} - \log\frac{\pi_\theta(\mathbf{y}_l \mid \mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}_l \mid \mathbf{x})}\right)\right)\right]$$

Understanding the limitation of DPO

A simple counter-example



A simple counter-example



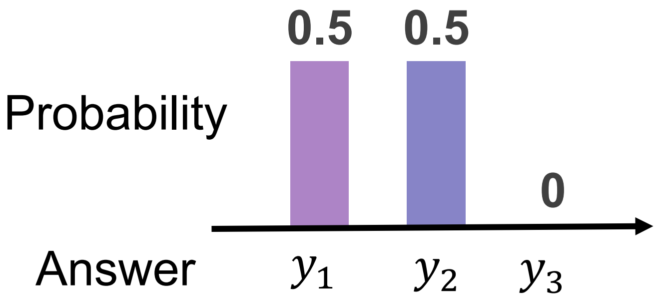
y_3 is an Out-of-Distribution answer

DPO fails to find the optimal policy…

WHY?

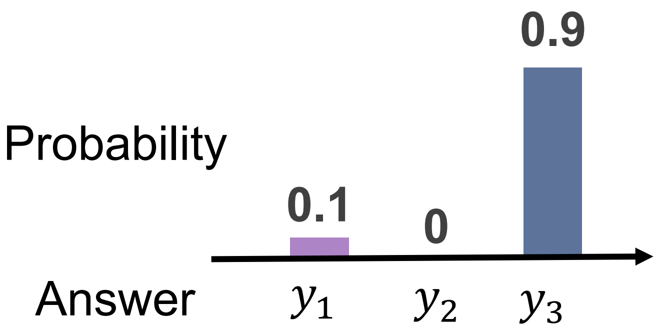
A simple counter-example

Reference Policy



DPO training

DPO Policy



+

Preference Dataset

y_1 ✅ y_2 ❌

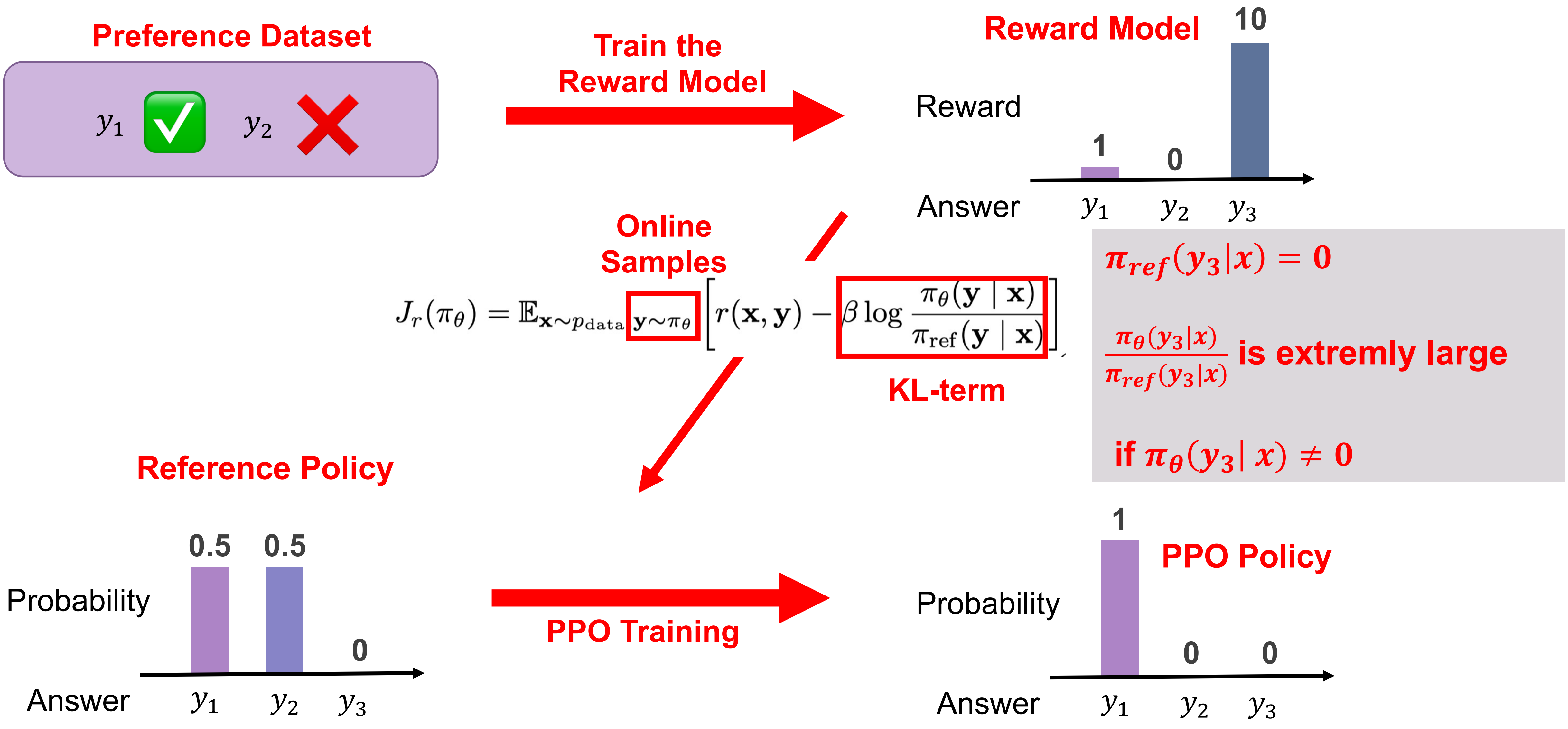
In this case

$$\mathcal{L}_{\text{DPO}}(\pi_\theta) = \log(1 + \left(\frac{\pi_\theta(y_2 \mid x)}{\pi_\theta(y_1 \mid x)}\right)^\beta) = 0$$

$\mathcal{L}_{DPO}$ is minimized when $\pi_\theta(y_2|x) = 0$, irrelevant to y_3 and y_1.

DPO is not guaranteed to find the optimal policy !

How does PPO work this case?



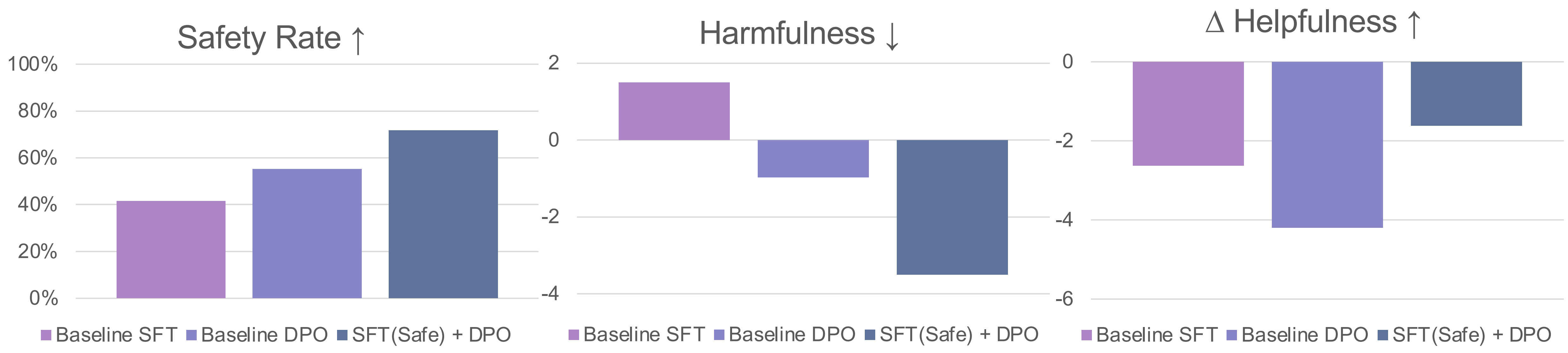
Preference Dataset
y_1 y_2
Train the Reward Model
Reward Model
Reward
10
1
0
Answer
y_1 y_2 y_3
Online Samples
$J_r(\pi_\theta) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}, \boxed{\mathbf{y}\sim\pi_\theta}}\left[r(\mathbf{x},\mathbf{y}) - \boxed{\beta \log \frac{\pi_\theta(\mathbf{y}\mid\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}\mid\mathbf{x})}}\right]$
KL-term
$\pi_{ref}(y_3|x) = 0$
$\frac{\pi_\theta(y_3|x)}{\pi_{ref}(y_3|x)}$ is extremly large
if $\pi_\theta(y_3|\, x) \neq 0$
Reference Policy
0.5
0.5
0
Probability
Answer
y_1 y_2 y_3
PPO Training
PPO Policy
1
0
0
Probability
Answer
y_1 y_2 y_3

How to improve DPO?

Understanding the Limitation of DPO

Experiments on the Real Preference Dataset: SafeRLHF[1]

solution 1: Additional SFT over the training dataset.



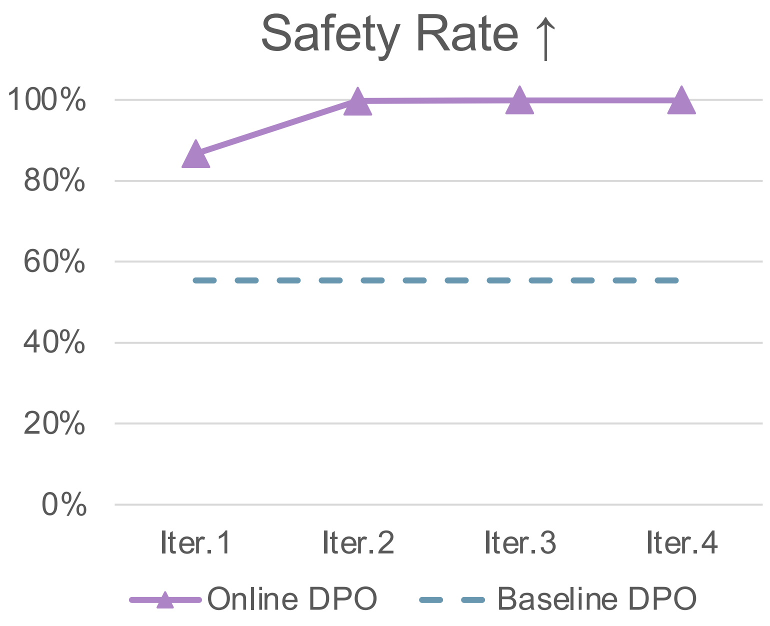
P.S. Helpfulness and safety are evaluated by the released model in the original paper.

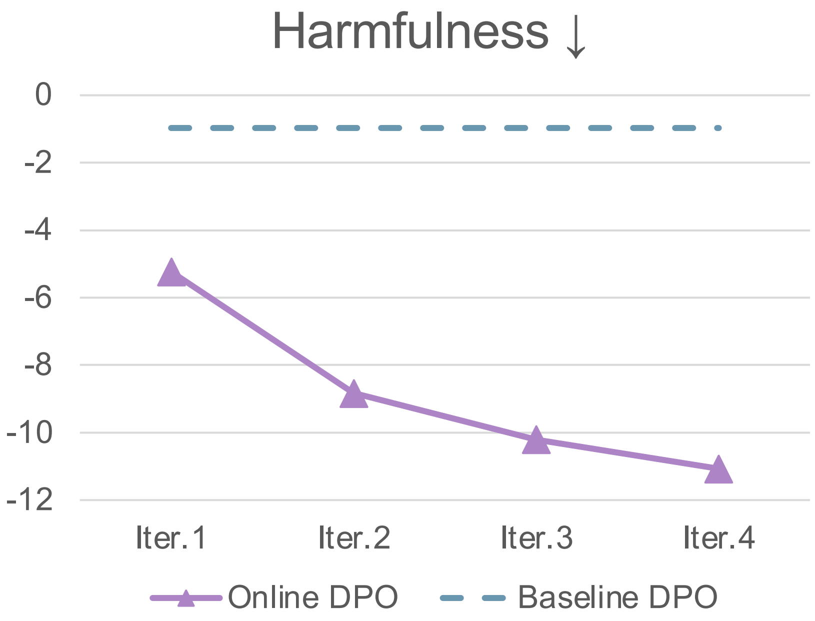
[1] Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., ... & Yang, Y. (2023). Safe RLHF: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773.*

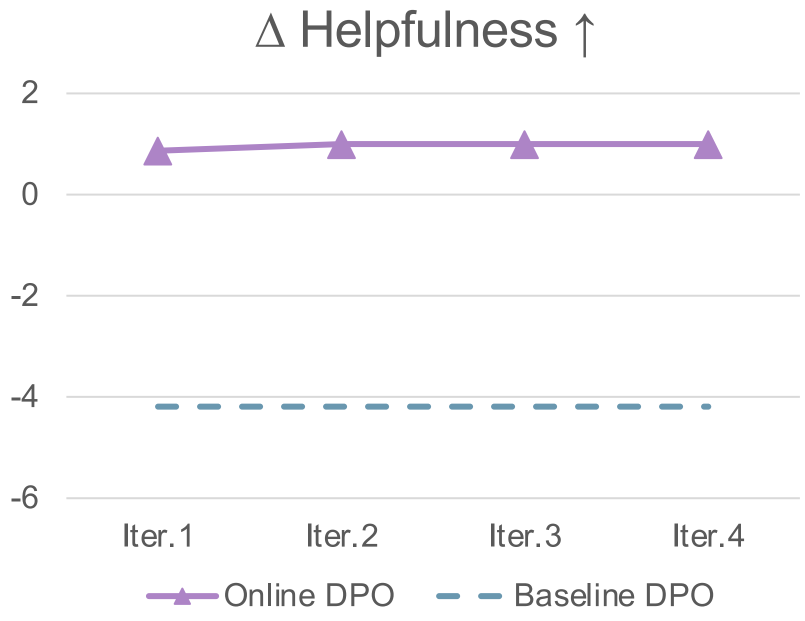
Understanding the Limitation of DPO

Experiments on the Real Preference Dataset: SafeRLHF[1]

solution 2: Online generation and scoring with a trained reward model.







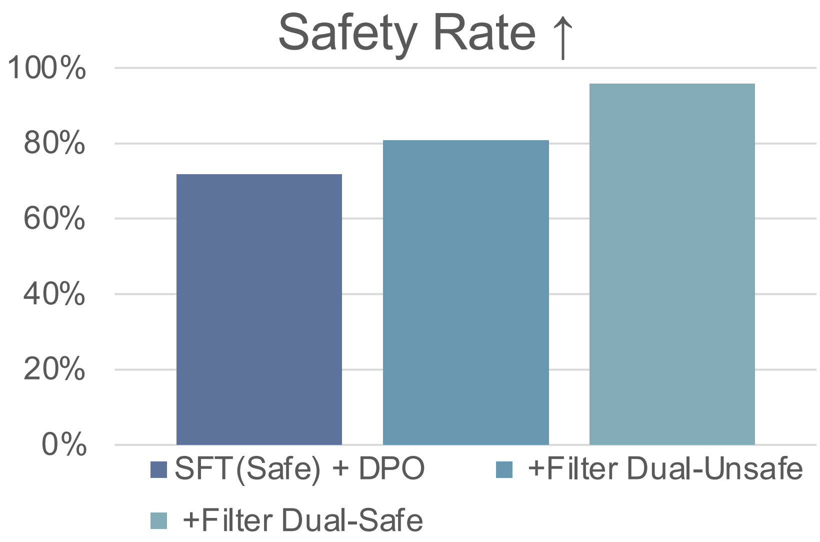
Helpfulness and safety are evaluated by the released model in the original paper.

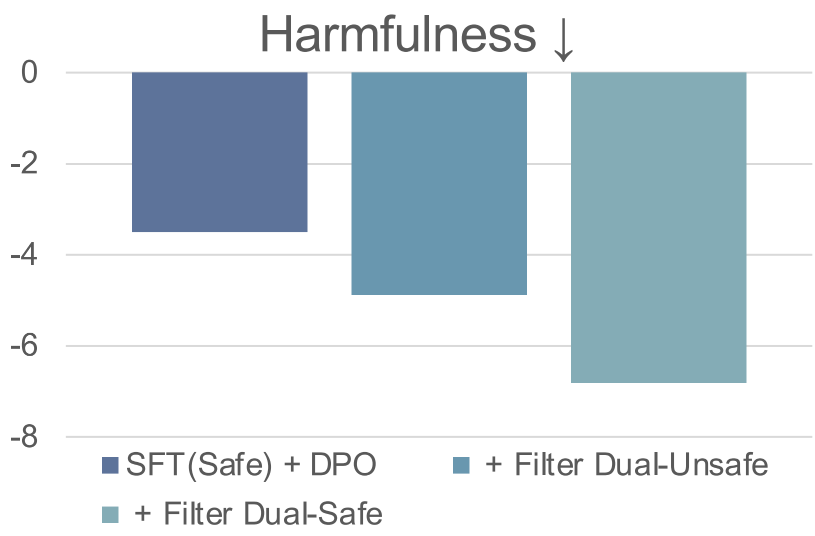
[1] Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., ... & Yang, Y. (2023). Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

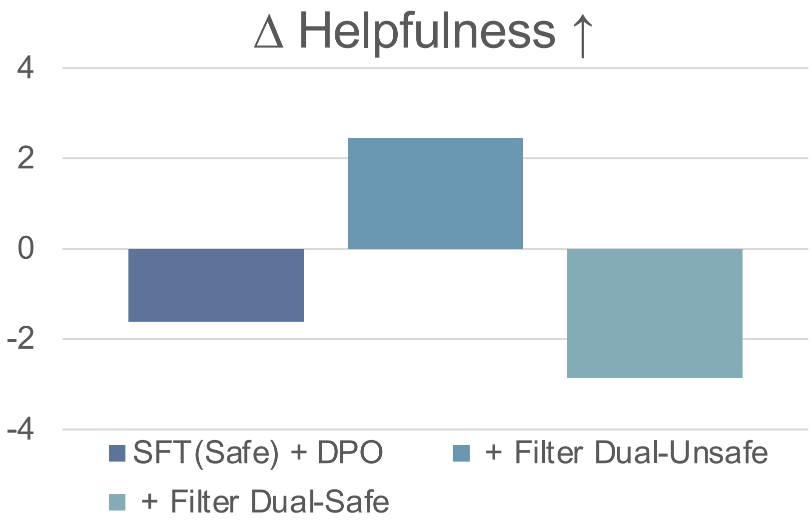
Understanding the Limitation of DPO

Experiments on the Real Preference Dataset: SafeRLHF[1]

Additional Trick: Eliminate noises or controversies in the dataset.







But this will filter out some high-quality data, thus hurt helpfulness!

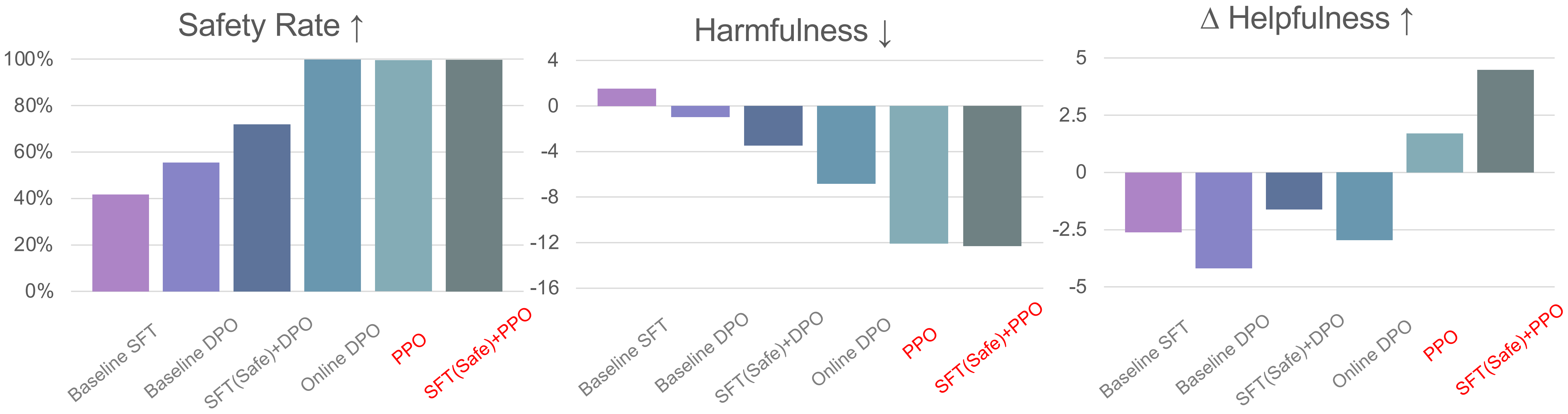
[1] Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., ... & Yang, Y. (2023). Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

What about PPO in the SafeRLHF benchmark?

Understanding the Limitation of DPO

Experiments on the Real Preference Dataset: SafeRLHF[1]

What about PPO in this benchmark? An end-to-end comparison with DPO.



[1] Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., ... & Yang, Y. (2023). Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

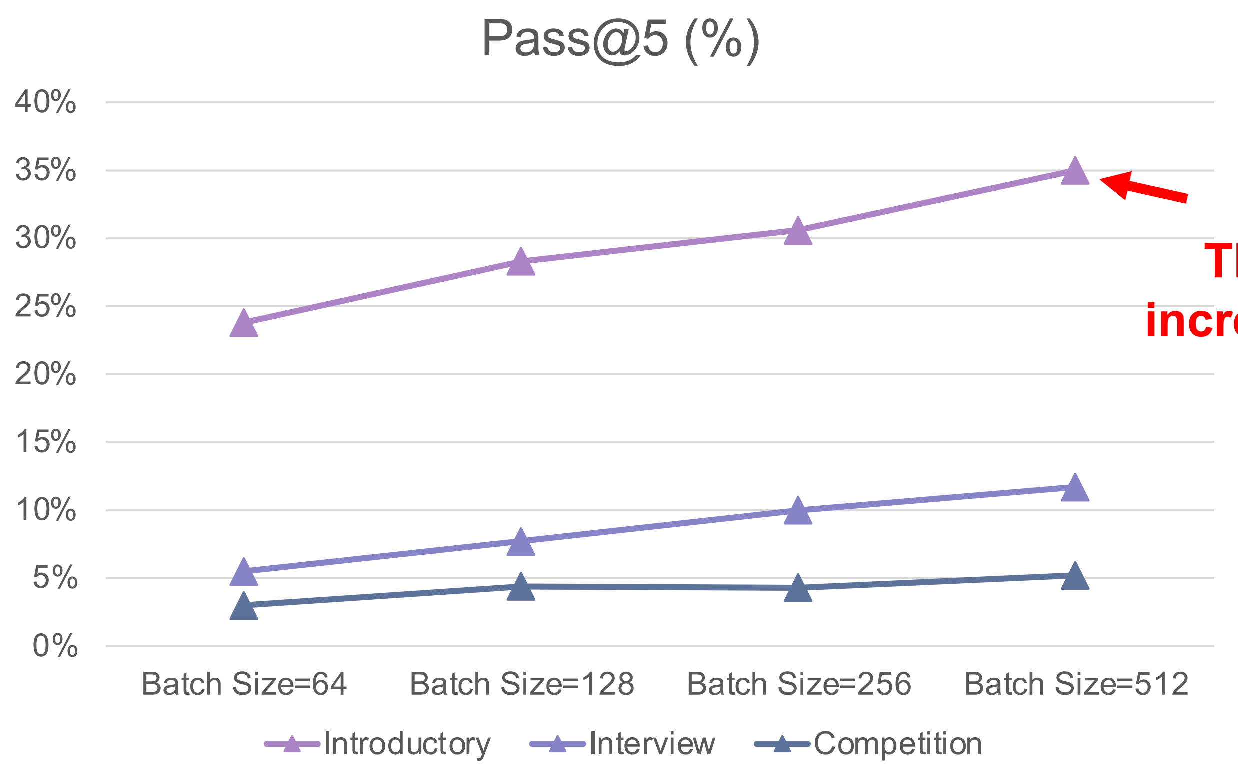
Key Factors to Improve the performance of PPO

1️⃣

*A **LARGE** Batch Size*

Key Factors to Improve the Performance of PPO

Competitive programming: APPS dataset



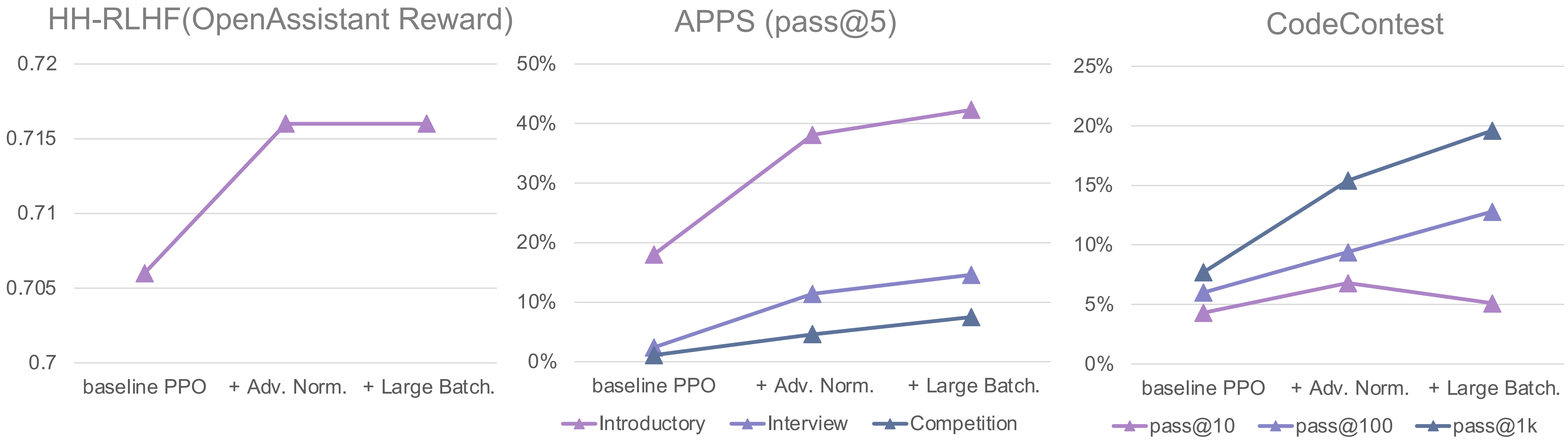
64→512:
The pass rate increases by ~10%

2️⃣

Advantage Normalization

Key Factors to Improve the Performance of PPO

Task: Competitive programming & conversation



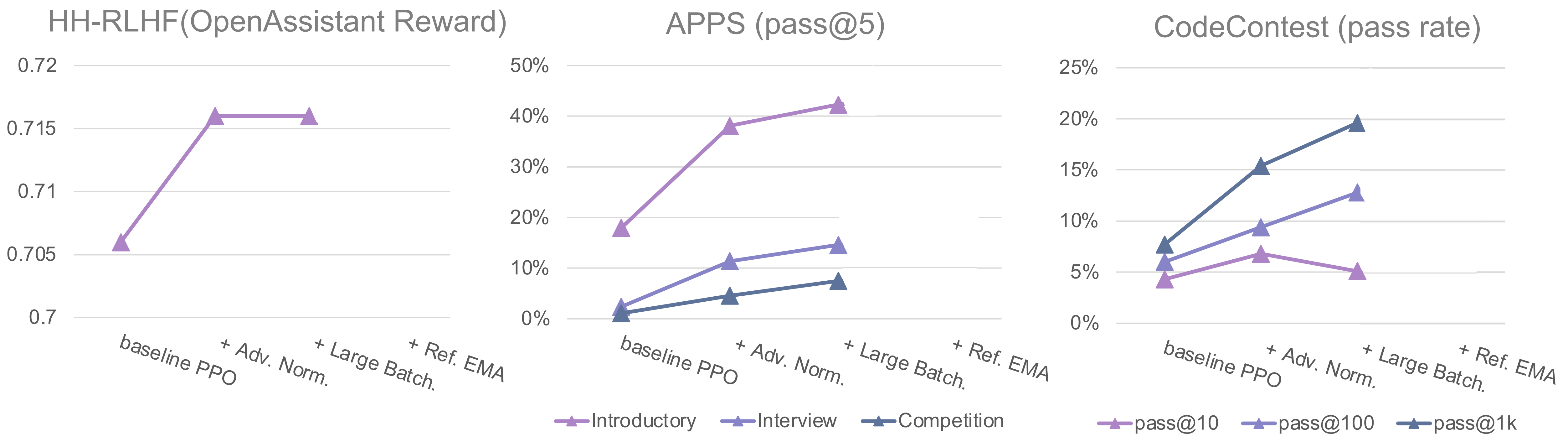
3️⃣

Exp. Moving Average for the Reference Model

Key Factors to Improve the Performance of PPO

Update the reference model with exponential moving average during training:

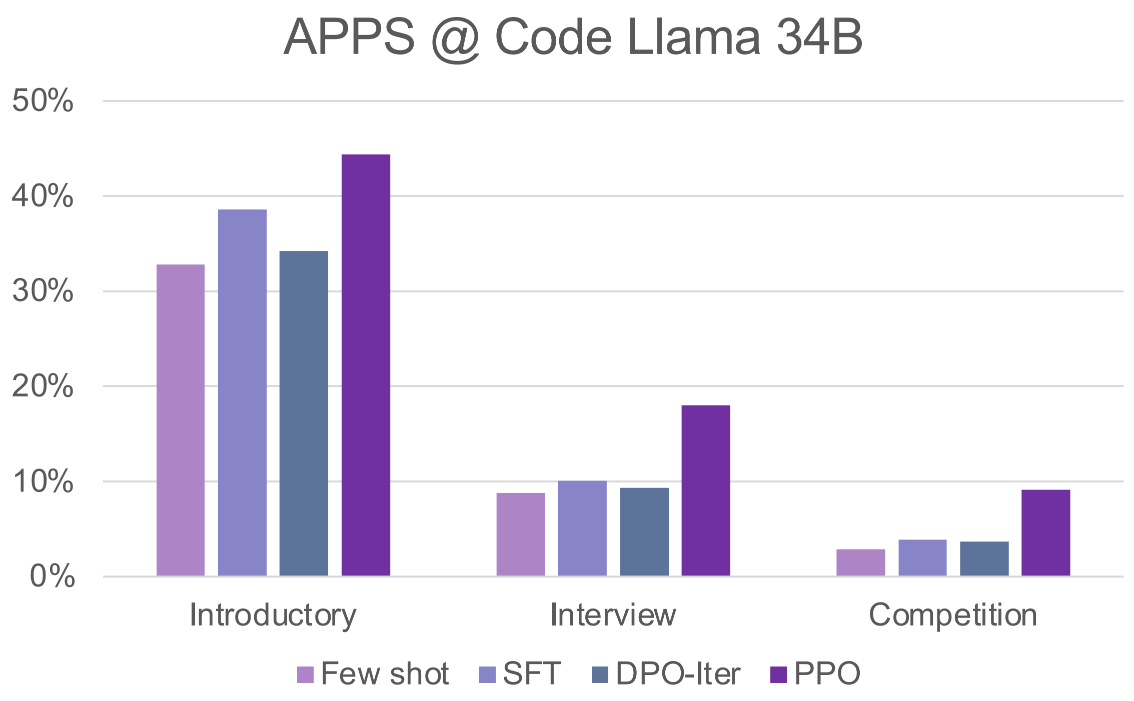
$$\pi_{\mathrm{ref},k} = \alpha\pi_{\mathrm{ref},k-1} + (1-\alpha)\pi_{\mathrm{actor},k}$$

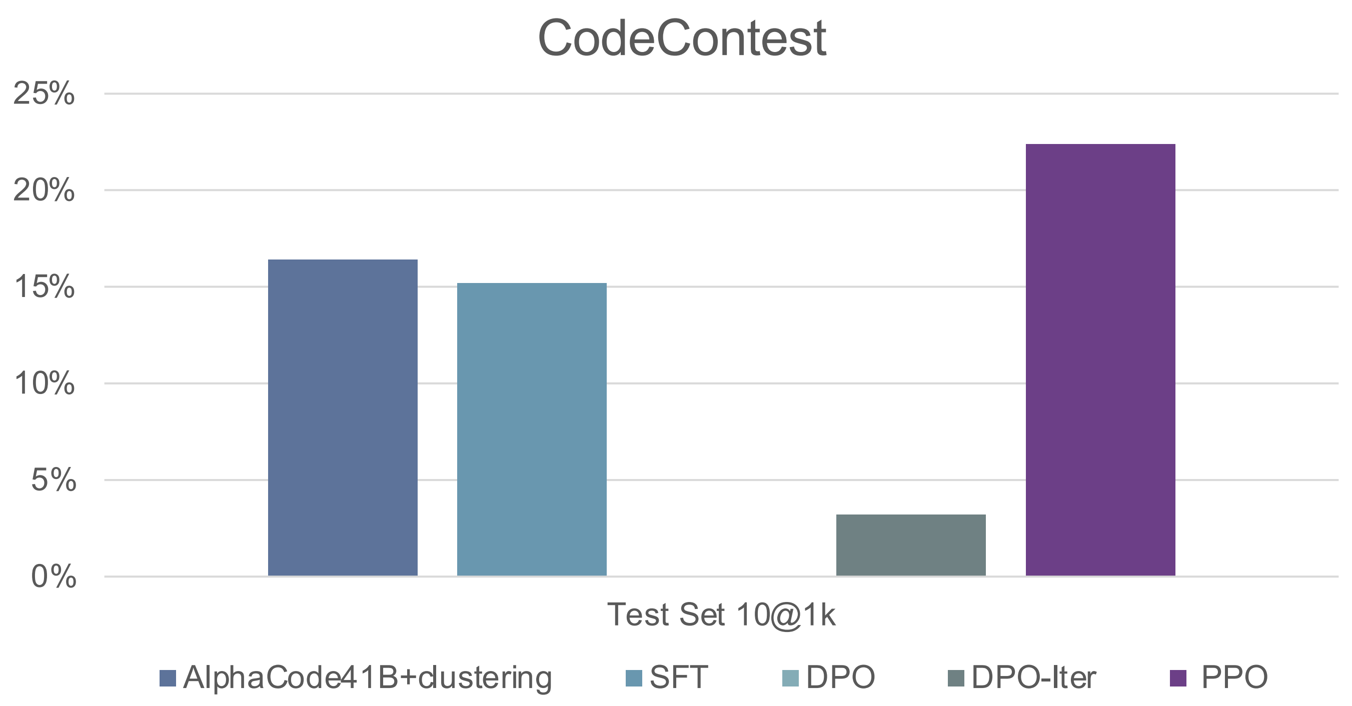


Benchmark Results

Benchmark Results

Task: Competitive Programming (test/validation set for APPS and CodeContest).

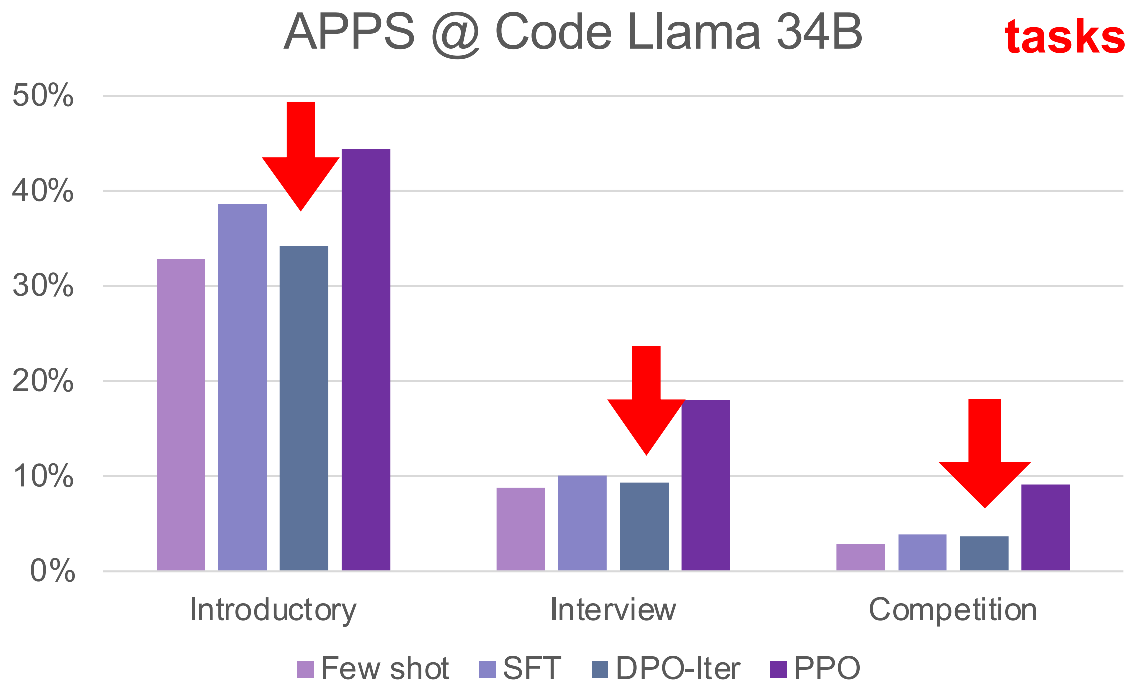


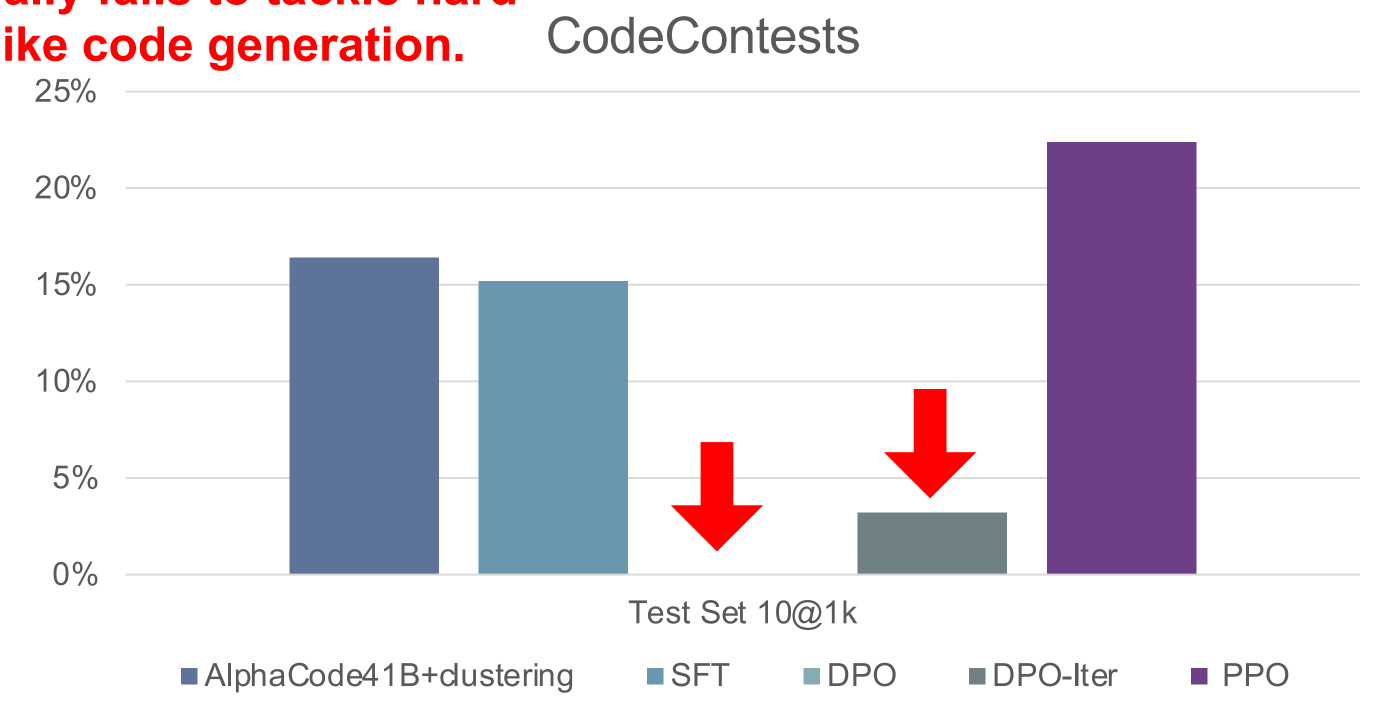


Benchmark Results

Task: Competitive Programming (test/validation set for APPS and CodeContests).

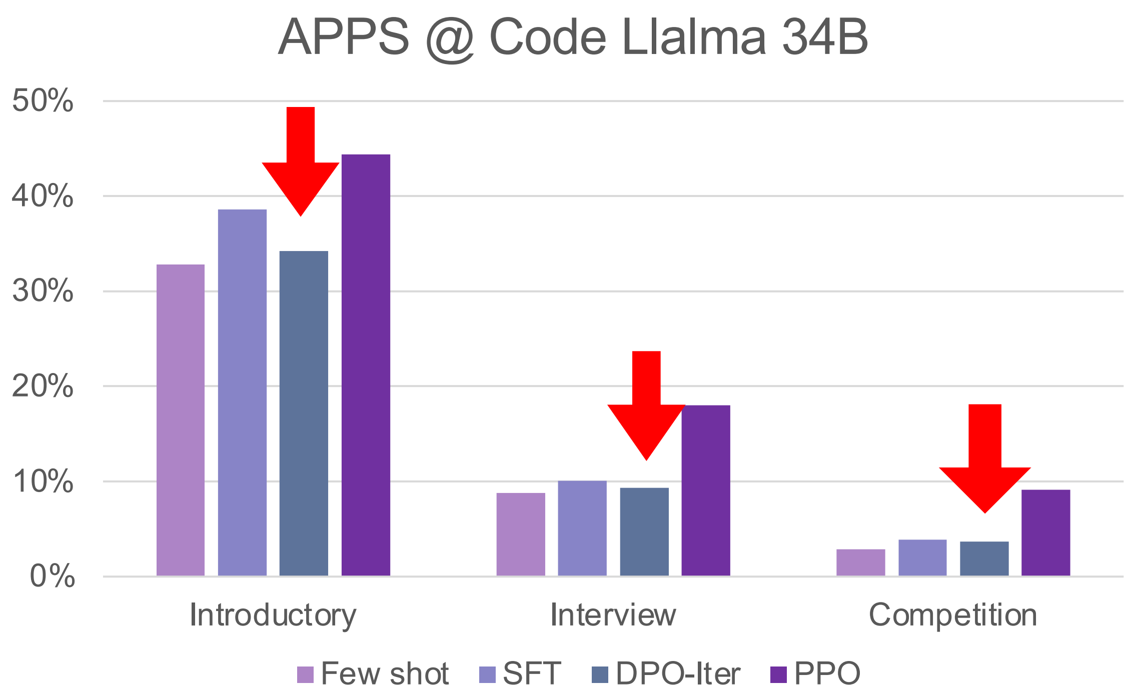
DPO usually fails to tackle hard tasks like code generation.

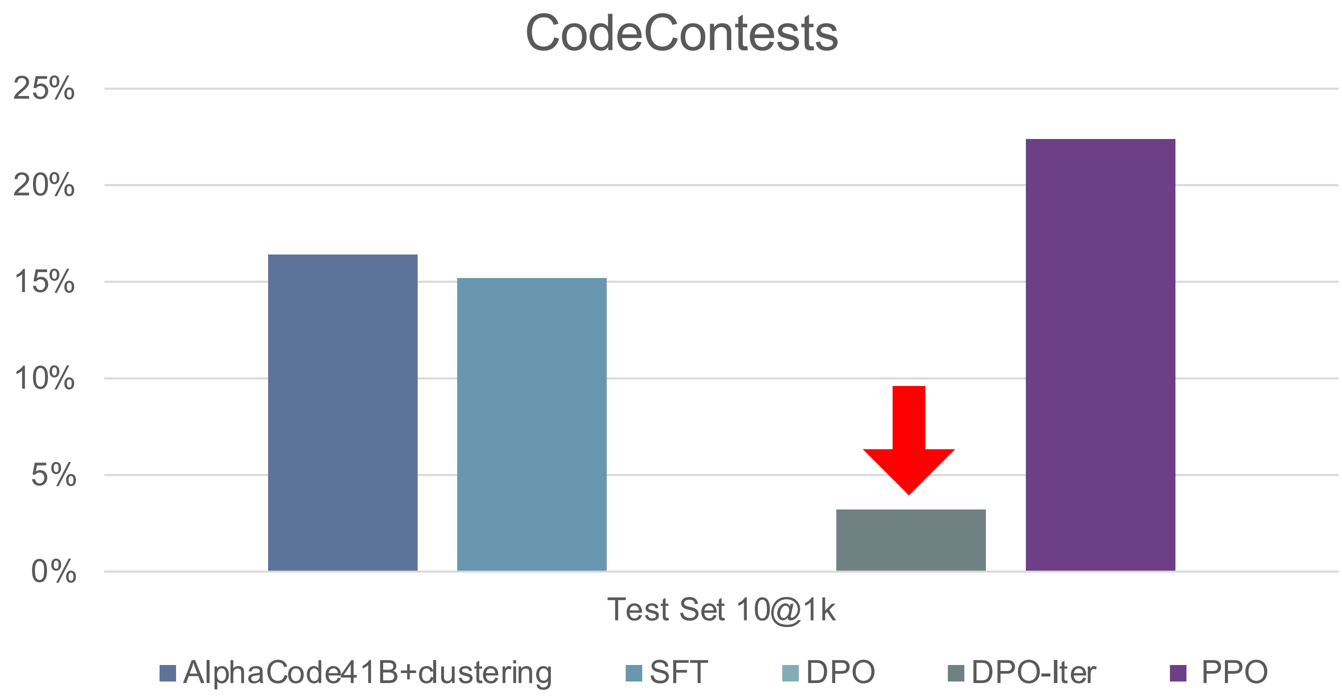




Benchmark Results

Task: Competitive Programming (test/validation set for APPS and CodeContests).

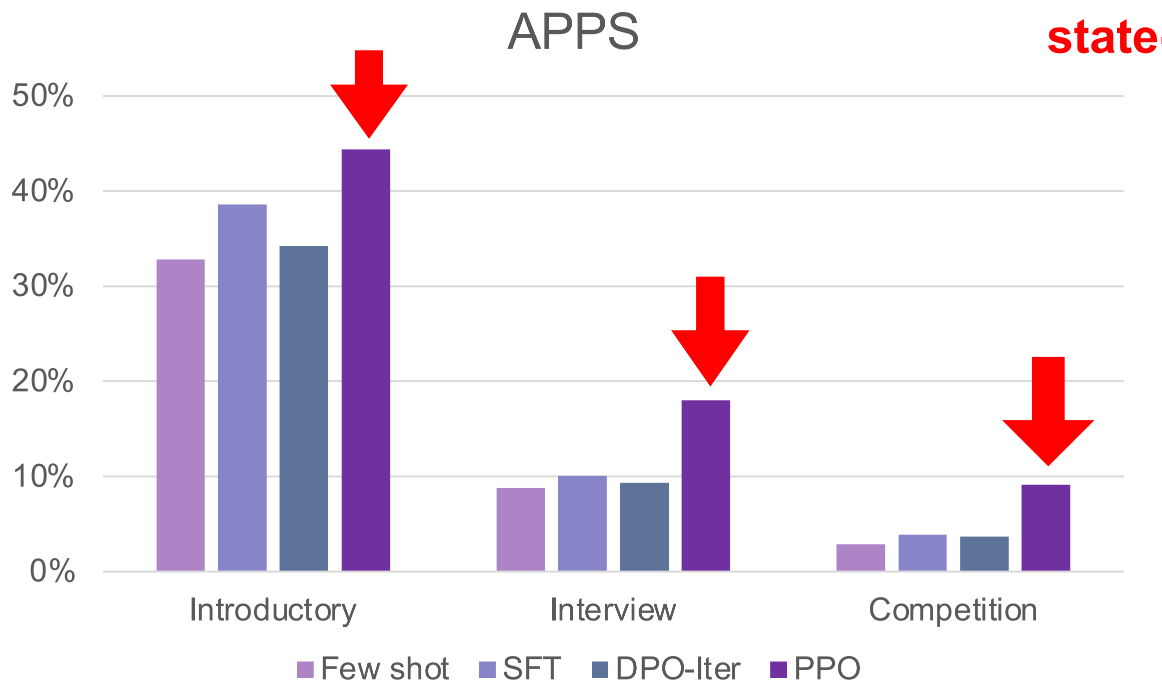


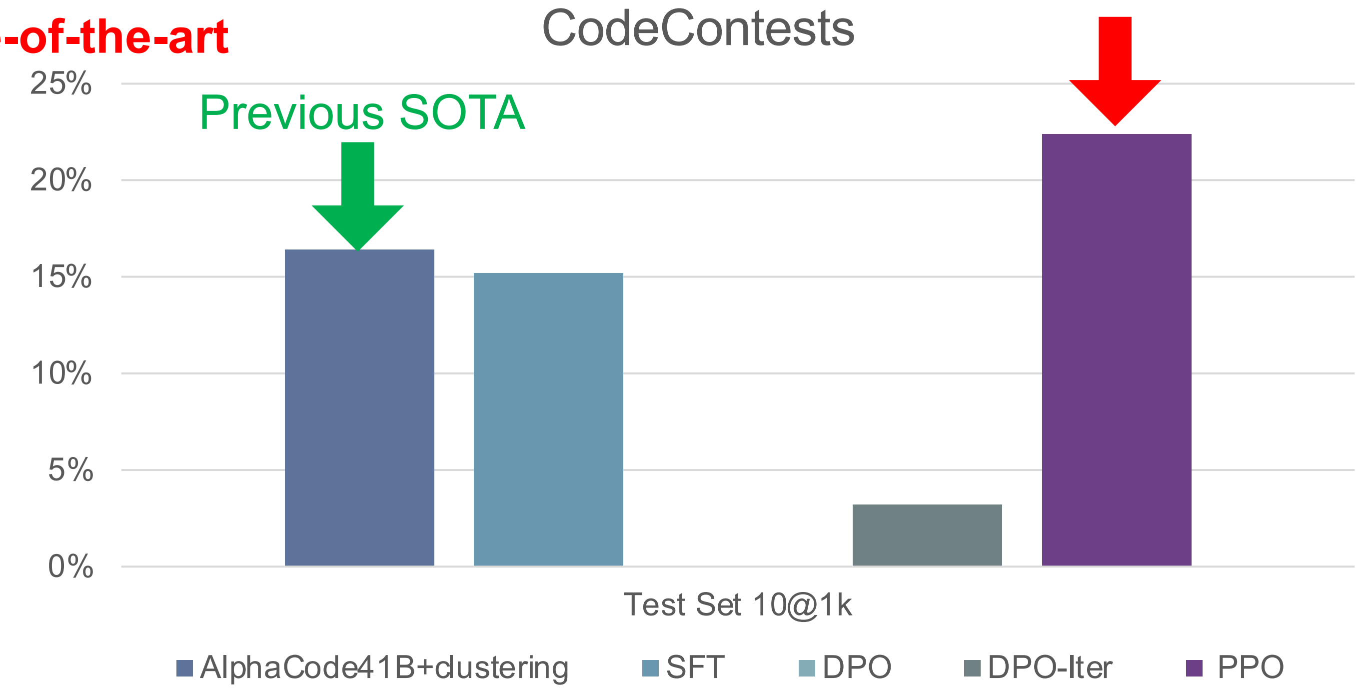


Benchmark Results

Task: Competitive Programming (test/validation set for APPS and CodeContests).

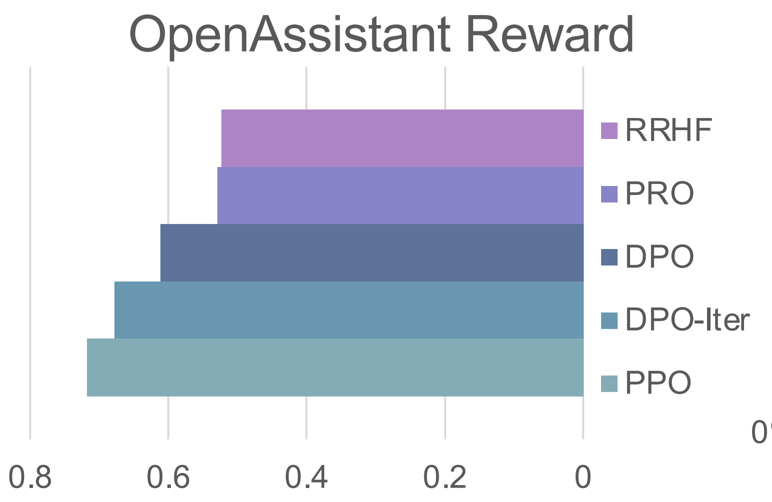
PPO establishes a new state-of-the-art

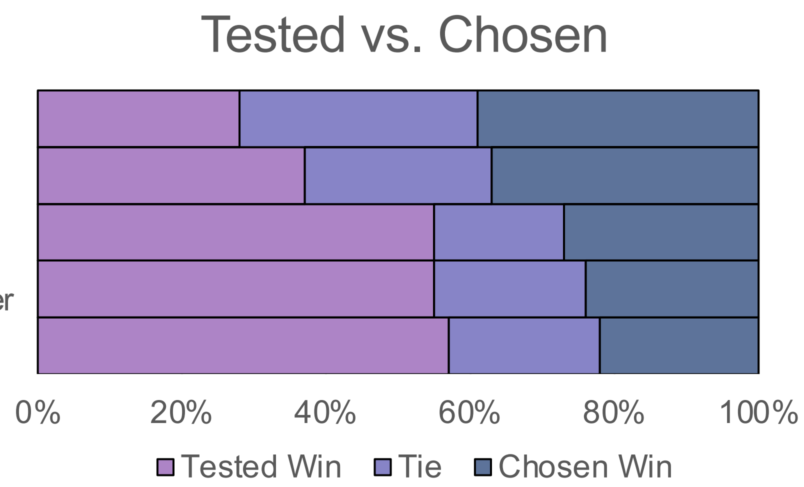


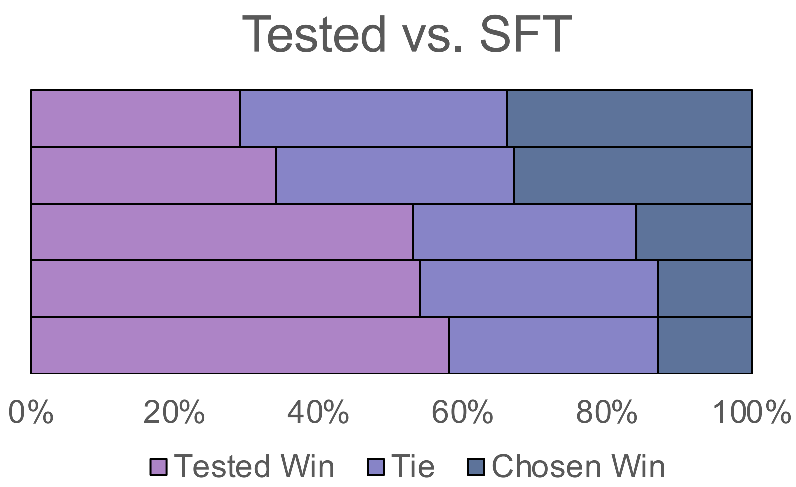


Is DPO Superior to PPO For LLM Alignment? A Comprehensive Study.

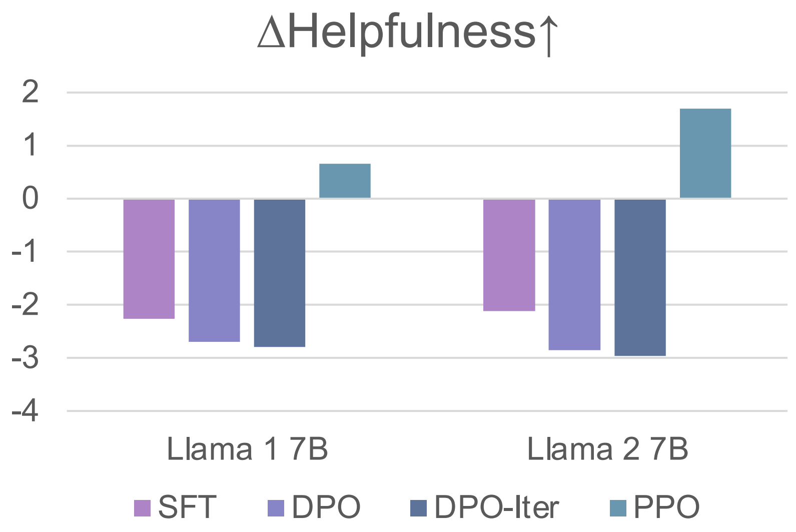
Task: HH-RLHF conversation.

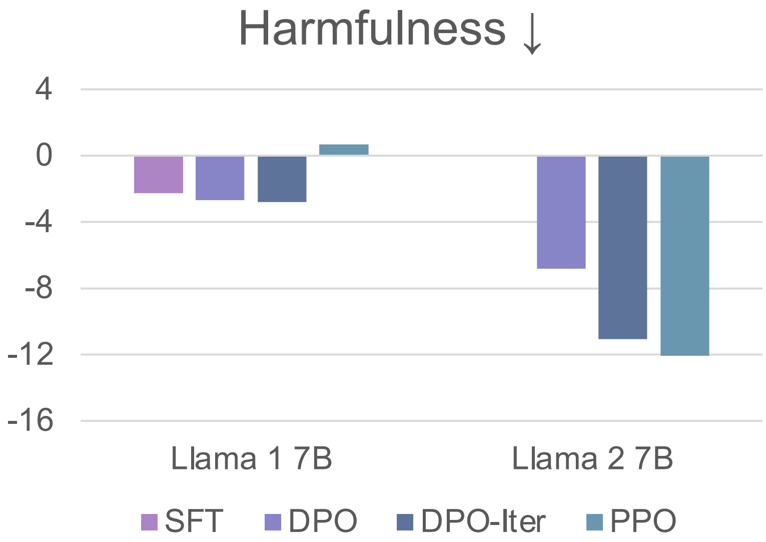






Task: SafeRLHF conversation.





Conclusion

Takeaways

- When applying DPO, we suggest
  - Performing an additional round of SFT over the accepted answers;
  - Carefully annotating data;
  - Iteratively generating fresh answers and labels for continuous learning.
- When applying PPO, we suggest using
  - A large batch size (512 sequences or larger),
  - Advantage normalization,
  - And exponential moving average of the reference model.

清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

上海期智研究院
SHANGHAI QI ZHI INSTITUTE

OpenPSI

Check our PPO code for training 70B LLMs at:
https://github.com/openpsi-project/ReaLHF!

👈 Or scan the QR code here.

Running PPO for 70B+ LLMs

清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

Thank you for listening!