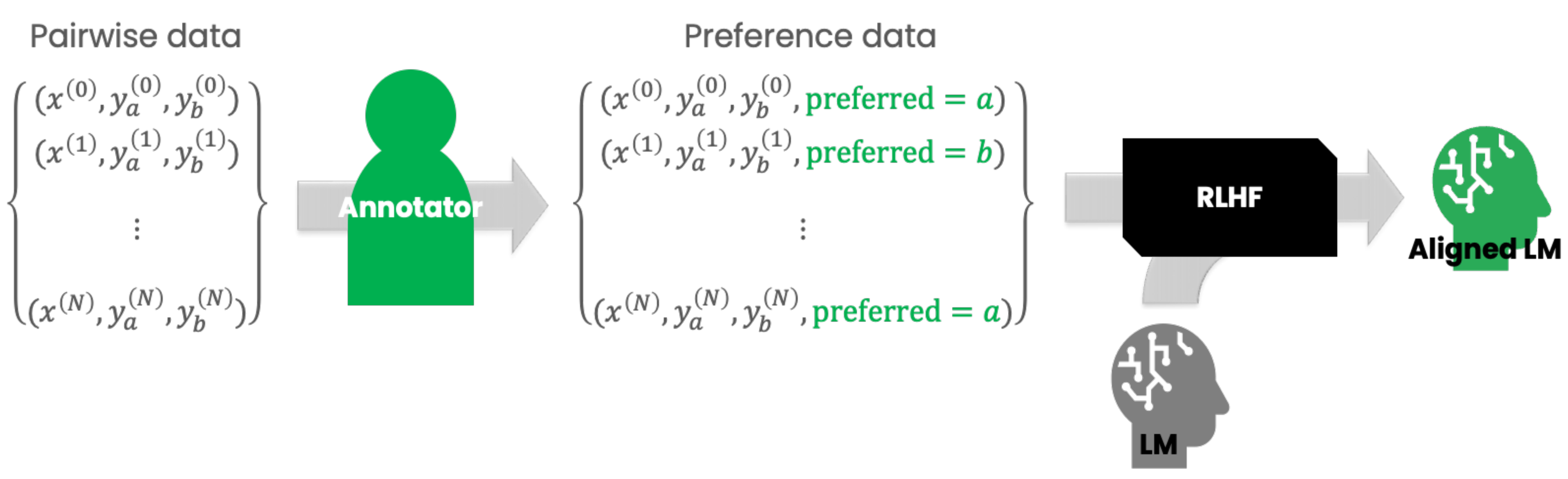
Filtered Direct Preference Optimization

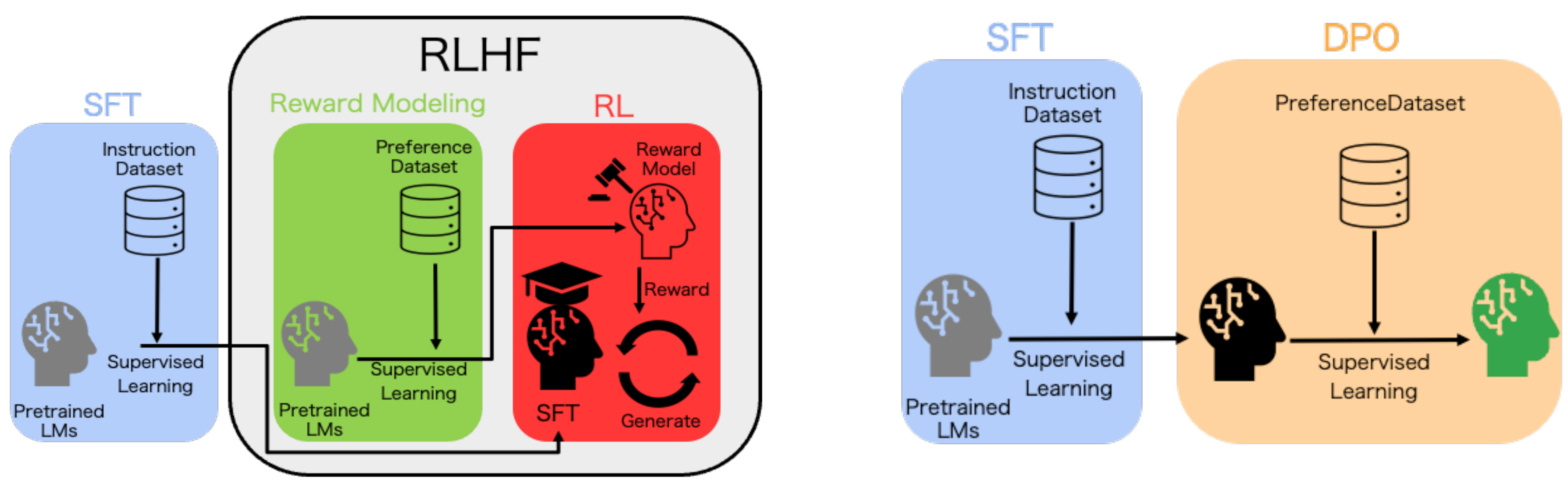
Tetsuro Morimura*, ○Mitsuki Sakamoto*, Yuu Jinnai, Kenshi Abe, Kaito Ariu
CyberAgent
*Equal Contribution
CyberAgent AI Lab

Introduction

- RLHF (Reinforcement Learning from Human Feedback) is essential for aligning language models (LMs) with human preferences
- Enhances practicality, trustworthiness, and social acceptance of LMs



Sensitivity of RLHF and DPO to data quality

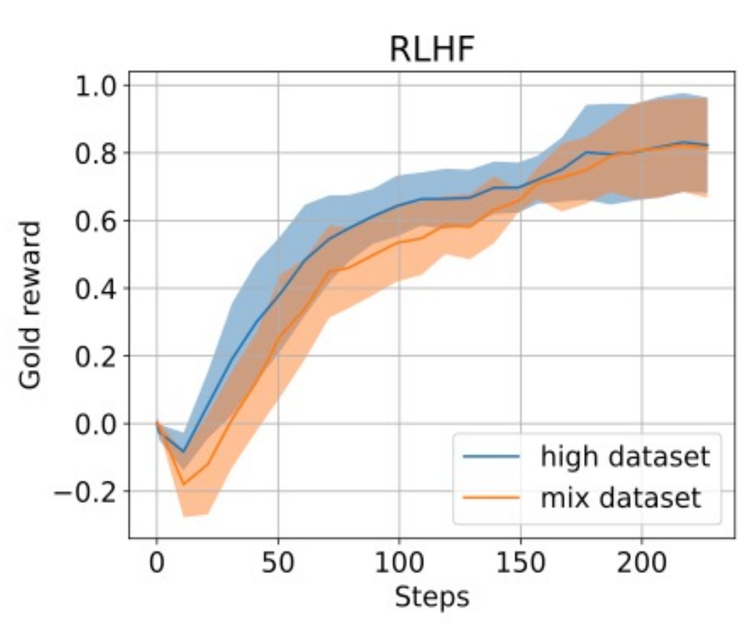


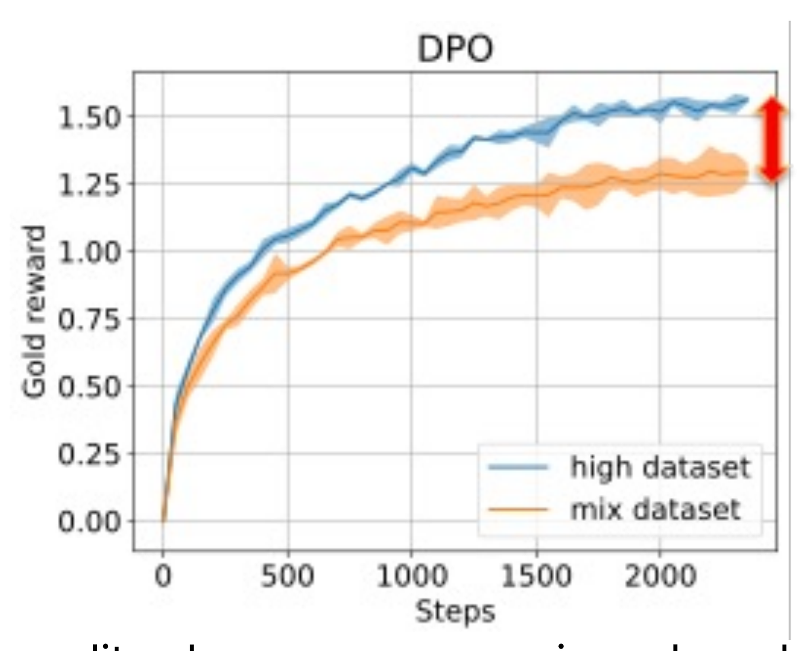
(RM-based) RLHF [Ouyang+ 2022]

- Uses a reward model (RM) to learn from preference data and optimize LM
- Data-efficient but computationally intensive

RM-free RLHF: DPO [Rafailov+ 2023]

- Directly optimizes the LM based on preference data without using an RM
- Simpler and less computationally intensive

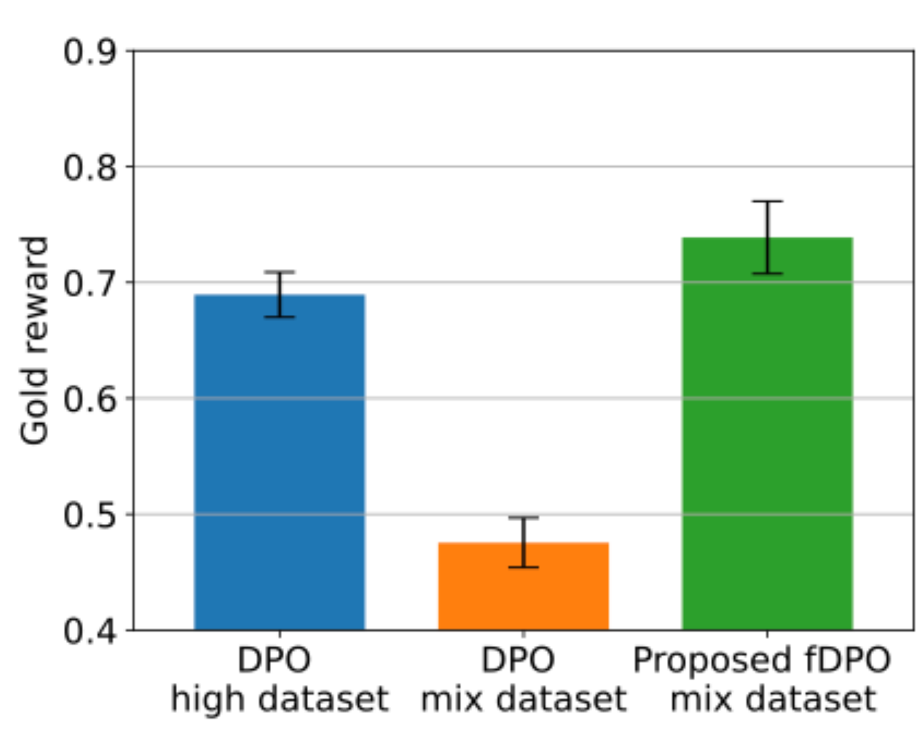




Mix-quality dataset: Half of the high-quality chosen responses is replaced with low-quality chosen responses . (5 independent runs)

Problem: Low-quality chosen responses are harmful for DPO

Experiment: Alpaca-Farm dataset [Dubois+ 2023]



- ✓ fDPO circumvented the performance decline observed with DPO
- ✓ Indicates its effectiveness in improving DPO performance where dataset quality is diverse

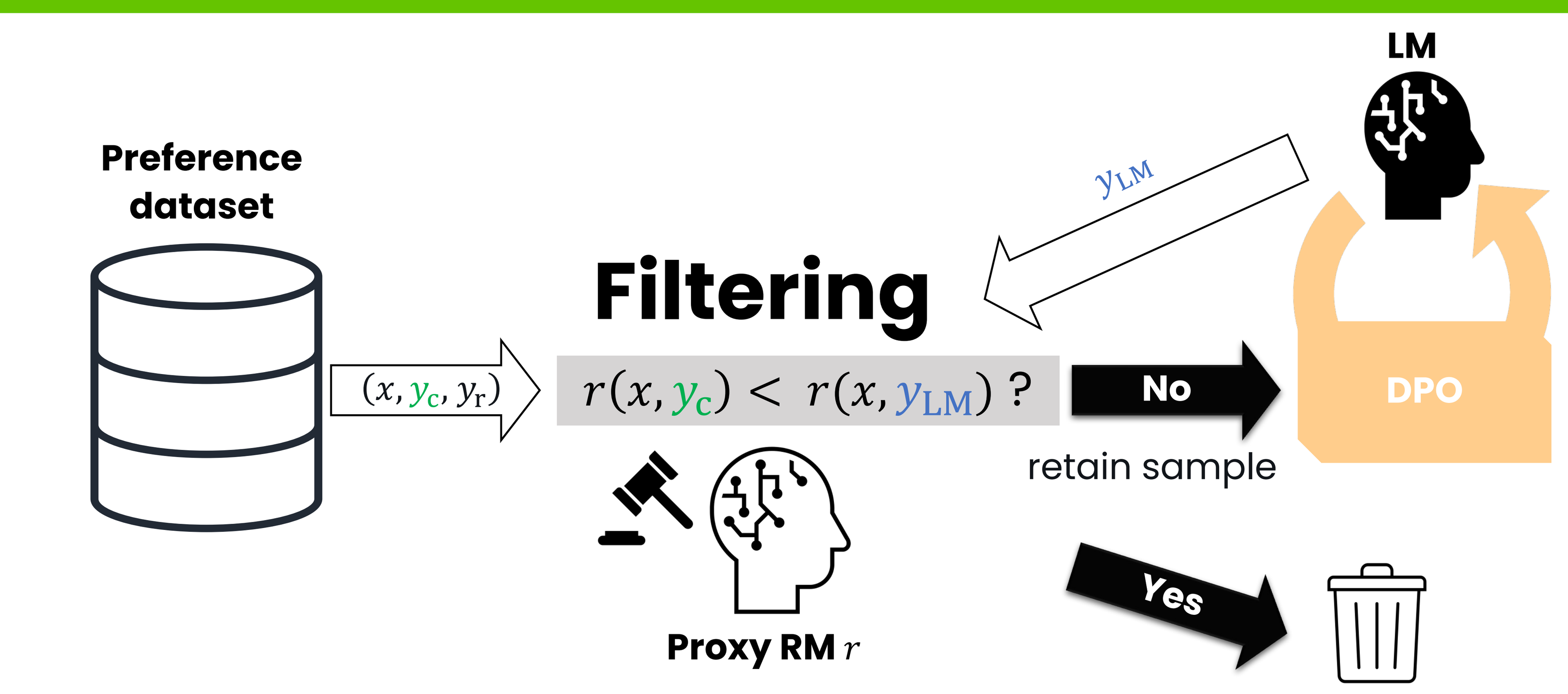
Gold rewards are adjusted so that the average reward of the SFT model is zero (5 independent runs)

- Details on varying quality
  - **Low-quality dataset**: 2 responses are generated by SFT and labeled by Gold RM. "Chosen" is assigned to the higher scoring response.
  - **High-quality dataset**: 16 responses are generated by SFT and labeled by Gold RM. "Chosen" is assigned to the highest scoring response, while "rejected" is assigned to one randomly selected from the remaining 15 responses.
  - **Mix-quality dataset**: Mixing the low-quality and high-quality datasets in a 50/50 ratio
- Models
  - **LLM** : 1.4B Pythia model [Biderman+ 2023]
  - **Proxy RM for filtering** : 160M Pythia model
    - This is trained on preference data.
  - **Gold RM**: OpenAssistant/reward-modeldeberta-v3-large-v2
    - This is used for annotation and evaluation, emulating human annotators.

Summary

- Quality of responses in datasets affects the performance of DPO
- fDPO uses a reward model to filter out low quality samples
- fDPO enhances LMs performance, when response quality varies

Proposed: Filtered DPO



Discard low quality chosen response

- Lower-quality chosen response can induce performance bottlenecks in DPO
- fDPO discards lower-quality chosen responses compared to those generated by the optimizing LM with Proxy RM

Algorithm 1 filtered direct preference optimization (fDPO)

Require: LM π_θ, RM r_ϕ, demonstration data $\mathcal{D}_{\text{demo}}$, preference data $\mathcal{D}_{\text{pref}}$, and maximum epoch M.

Step 1: Supervised fine-tuning. Train π_θ on $\mathcal{D}_{\text{demo}}$.
Step 2: Reward modeling. Train r_ϕ on $\mathcal{D}_{\text{pref}}$ (see Eq. (1)).
Step 3: DPO fine-tuning with filtering.
Initialize filtered-preference data $\mathcal{D}_{\text{filtered}} := \mathcal{D}_{\text{pref}}$, epoch number $m := 0$.
while $m < M$ **do**
  for each (x, y_c, y_r) in $\mathcal{D}_{\text{pref}}$ **do**
    Generate response y by LM π_θ given prompt x.
    if $r_\phi(x, y) > r_\phi(x, y_c)$ **then**
      Discard (x, y_c, y_r) from $\mathcal{D}_{\text{filtered}}$.
    end if
  end for
  Update preference data $\mathcal{D}_{\text{pref}} := \mathcal{D}_{\text{filtered}}$
  Update LM π_θ on $\mathcal{D}_{\text{pref}}$ for one epoch using DPO.
  Increment epoch number $m := m + 1$.
end while
return Optimized LM π_θ.

Additional procedure to DPO: Discards sample of lower quality chosen responses than LM-generated responses

Experiment: Realistic RLHF settings on Anthropic HH datasets [Bai+ 2022]

Dataset	Method	Gold RM Score (SFT=0.0) ↑	GPT-4o Evaluation (win rate vs. SFT) ↑
Helpful	DPO	1.42 ± 0.08	0.543 ± 0.015
	fDPO	$\mathbf{1.94 \pm 0.02}$	$\mathbf{0.628 \pm 0.001}$
Harmless	DPO	2.66 ± 0.12	0.891 ± 0.003
	fDPO	$\mathbf{3.20 \pm 0.06}$	$\mathbf{0.944 \pm 0.005}$

(3 independent runs)

- ✓ Indicates the effectiveness of fDPO under realistic RLHF settings
- ✓ Superior GPT-4o results suggest higher-quality human-like responses

- Realistic RLHF settings
  - Considered a realistic RLHF setting where **the number of high-quality responses created by humans is limited**
  - Instead of generating responses manually,
    - SFT created response pairs
    - Gold RM annotator provided labels (chosen or rejected) to the pairs
  - Created **mix-quality datasets by combining the original dataset and generated SFT samples**
- Models
  - **LLM**: 2.8B Pythia model
  - **Proxy RM for filtering** : 160M Pythia model
  - **Gold RM**: OpenAssistant/reward-modeldeberta-v3-large-v2

Future directions

- Leverage publicly available high-quality reward models within fDPO for improved performance
- Incorporate rejected responses in the filtering process to further enhance fDPO effectiveness
- Explore combining fDPO with other DPO-related extensions and conduct comparisons with other RLHF methods, particularly with larger LM