Variants of Reinforcement Learning with Human Feedback

Pu Yang

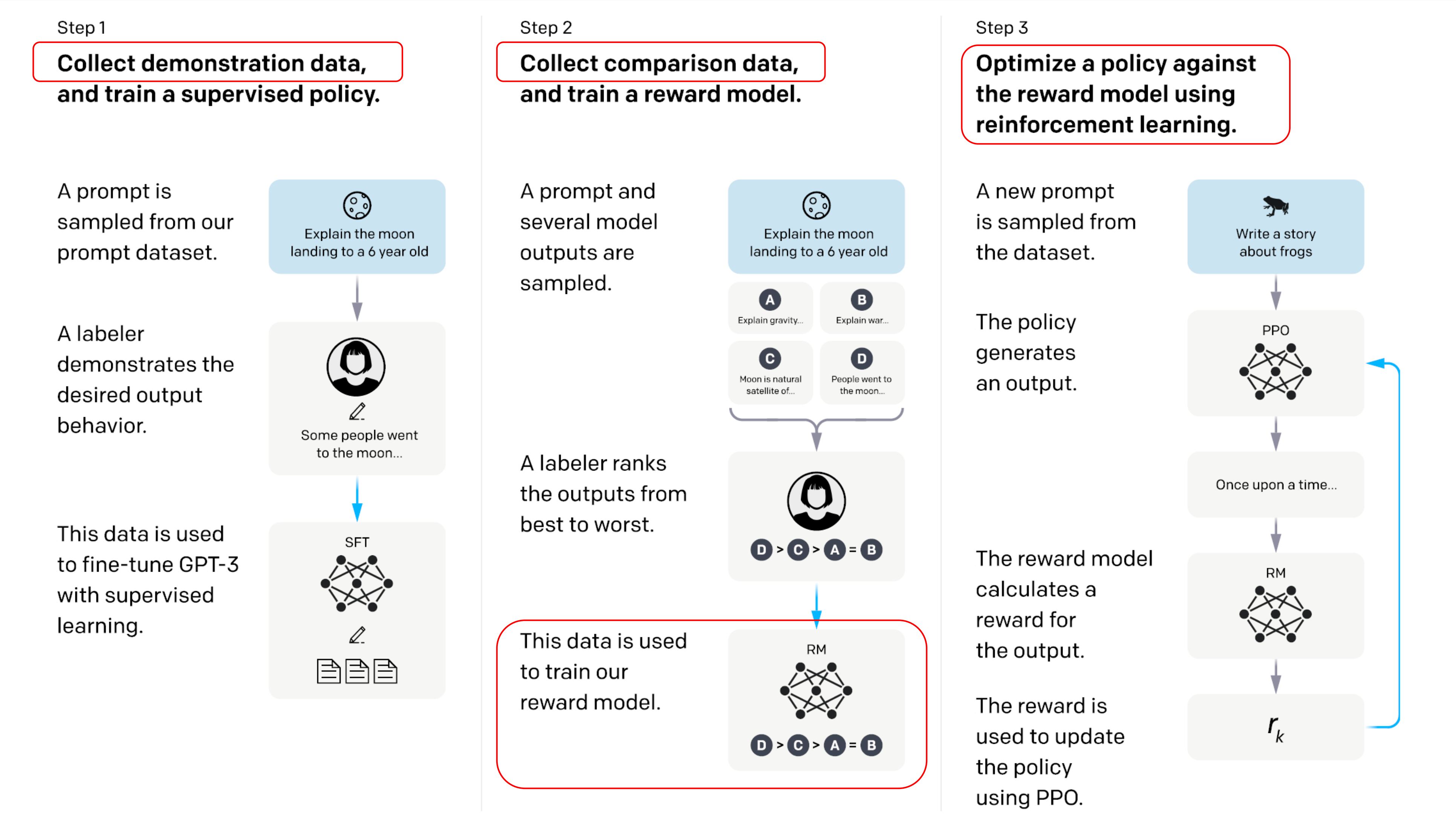
2024.3.20

Content

- **Review of Reinforcement Learning with Human Feedback (RLHF)**
- **Offline Methods**
  - Direct Preference Optimization (DPO)
  - Statistical Rejection Sampling Improves Preference Optimization (RSO)
- **Online Methods**
  - Reinforced Self-Training (ReST)
- **Discussion**

Review of Reinforcement Learning with Human Feedback (RLHF)

Review of Reinforcement Learning with Human Feedback



Preliminaries

- Supervised Fine-Tuning (SFT) Phase
- Reward Modelling Phase

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}.$$

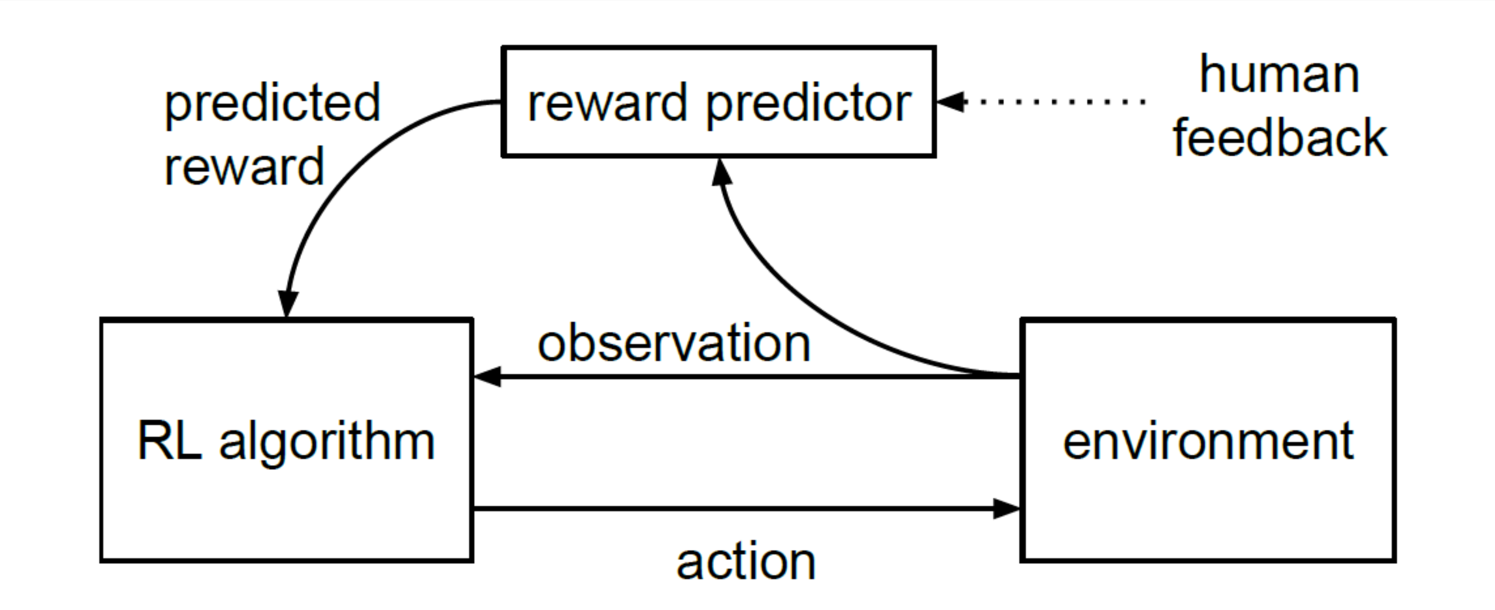
$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right]$$

- Reinforcement Learning (RL) Fine-Tuning Phase

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)}\left[r_\phi(x, y)\right] - \beta \mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \mid\mid \pi_{\mathrm{ref}}(y \mid x)\right]$$

Overview

- Inaccessible to the complete environment
- Model-based method:



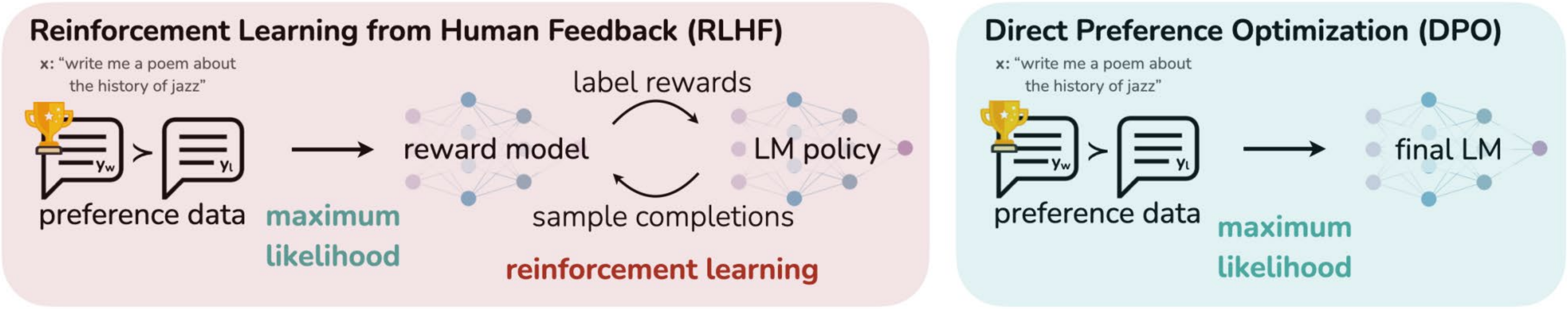
Offline Method:

Direct Preference Optimization (DPO)

Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." *Advances in Neural Information Processing Systems* 36 (2024).

Overview

- DPO optimizes for human preferences while avoiding reinforcement learning.



Optimal Solution to the KL-constrained Reward Maximization

$$\begin{aligned}\max_{\pi} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi}\big[r(x,y)\big] - \beta\mathbb{D}_{\mathrm{KL}}\big[\pi(y|x)\;||\;\pi_{\mathrm{ref}}(y|x)\big] \\ &= \max_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[r(x,y) - \beta\log\frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right] \\ &= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)} - \frac{1}{\beta}r(x,y)\right] \\ &= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)} - \log Z(x)\right]\end{aligned}$$

where we have partition function:

$$Z(x) = \sum_{y}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right).$$

Optimal Solution to the KL-constrained Reward Maximization

- We define

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right),$$

- Then re-organize the objective function as

$$\min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \left[\mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi^*(y|x)} \right] - \log Z(x) \right] =$$
$$\min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \left[\mathbb{D}_{\text{KL}}(\pi(y|x) \mid\mid \pi^*(y|x)) - \log Z(x) \right]$$

Optimal Solution to the KL-constrained Reward Maximization

- Gibbs' inequality tells us that the KL-divergence is minimized at 0 if and only if the two distributions are identical. Hence, we have the optimal solution

$$\pi(y|x) = \pi^*(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

- However, it is still expensive to estimate the partition function Z(x).

Re-organize Preference Model

- We can express the (unavailable) ground-truth reward as:

$$r^*(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)$$

- Then we obtain:

$$\begin{aligned}
p^*(y_1 \succ y_2|x) &= \frac{\exp\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} + \beta \log Z(x)\right)}{\exp\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} + \beta \log Z(x)\right) + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} + \beta \log Z(x)\right)} \\
&= \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)} \\
&= \sigma\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} - \beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right).
\end{aligned}$$

DPO Objective Function

- Analogous to the reward modeling approach, we can formulate a maximum likelihood objective for a parametrized policy π_θ.

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right].$$

- What does the DPO update do?

The gradient with respect to the parameters θ can be written as:

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$
$$-\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \bigg[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \bigg[\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease likelihood of } y_l} \bigg] \bigg],$$

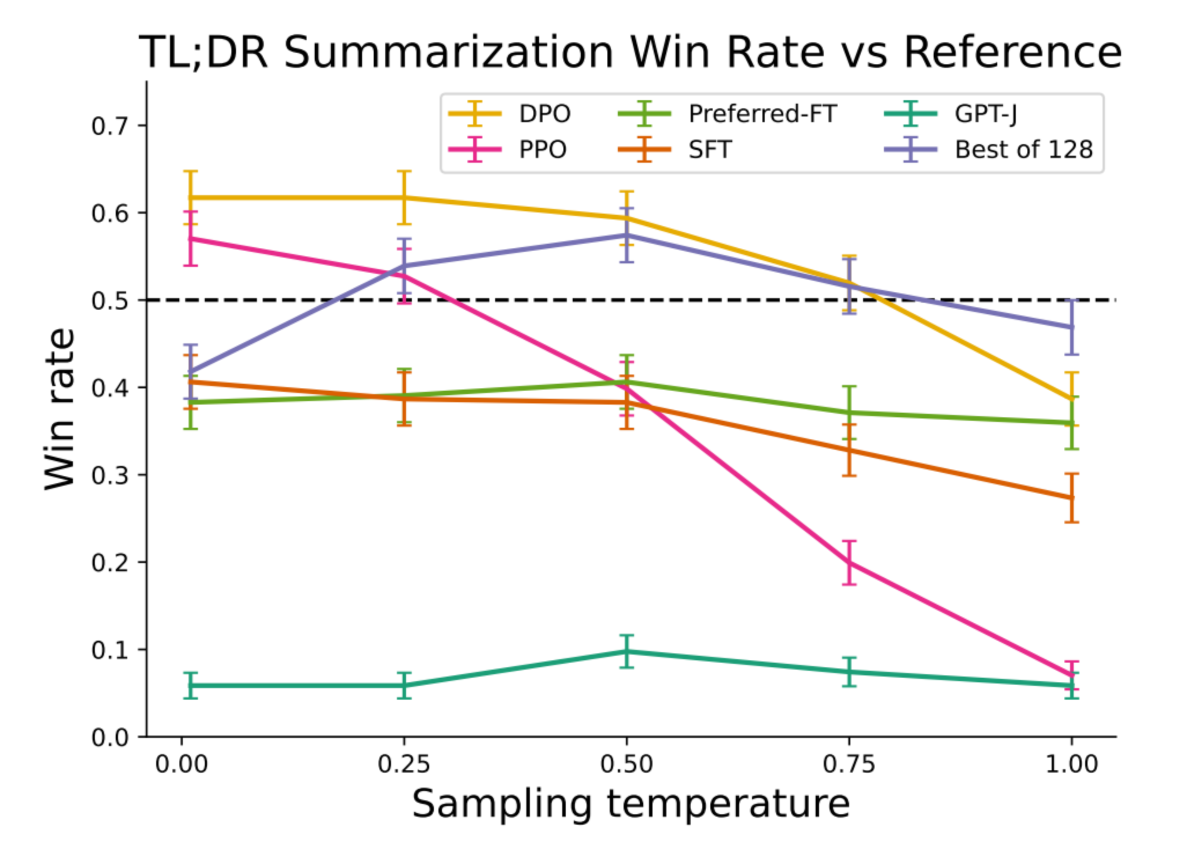
where $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ is the implicit reward.

DPO Outline

- 1. Sample completions $y_1, y_2 \sim \pi_{ref}(\cdot \mid x)$ for every prompt x, label with human preferences to construct the offline dataset of preferences $D = \{(x, y_w, y_l)\}$.
- 2. optimize the language model π_θ to minimize the DPO loss for the given π_{ref} and D and desired β.

Compared to PPO

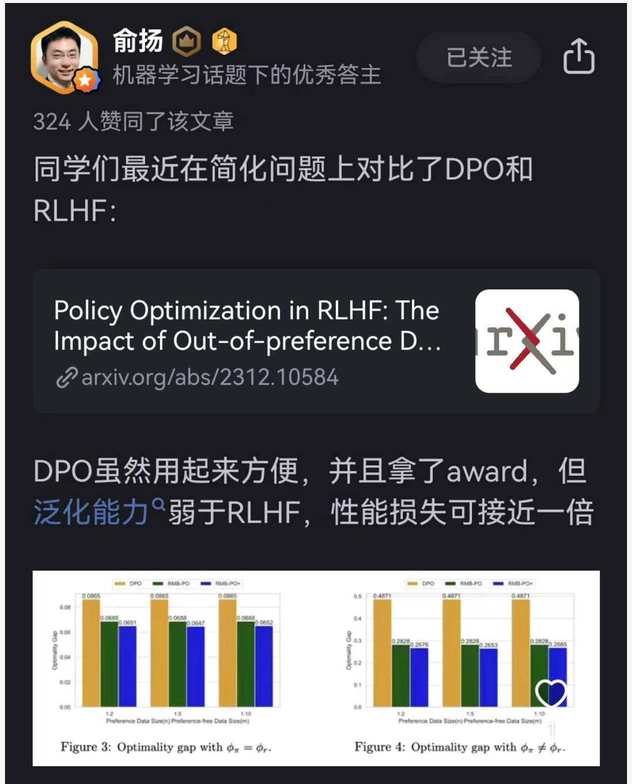
- Task: x is a forum post from Reddit; the policy must generate a summary y of the main points in the post.



Conclusion

- Advantage: 训练方便（不需要一边inference一边train），节省GPU memory

- Disadvantage: 效果褒贬不一



Offline Method:

Statistical Rejection Sampling Improves Preference Optimization (RSO)

Liu, Tianqi, et al. "Statistical Rejection Sampling Improves Preference Optimization." *The Twelfth International Conference on Learning Representations*. 2023.

Preference Data Distribution -- Intuition

- Suppose we have access to the oracle preference data

  $D^\star = \{(x, y_w, y_l) \mid y_w, y_l \sim \pi^\star(y \mid x)\}$, we can directly fit an MLE on the dataset.

- In reality, we have access to $\mathrm{D}_{\mathrm{hf}} = \{(x, y_w, y_l) \mid y_w, y_l \sim \pi_{unk}(y \mid x)\}$, where π_{unk} denotes some mixed unknown policies. The mixed unknown policies can include SFT policy, previous or current RLHF policy, or policies from other agents.

Preference Data Distribution -- Choices

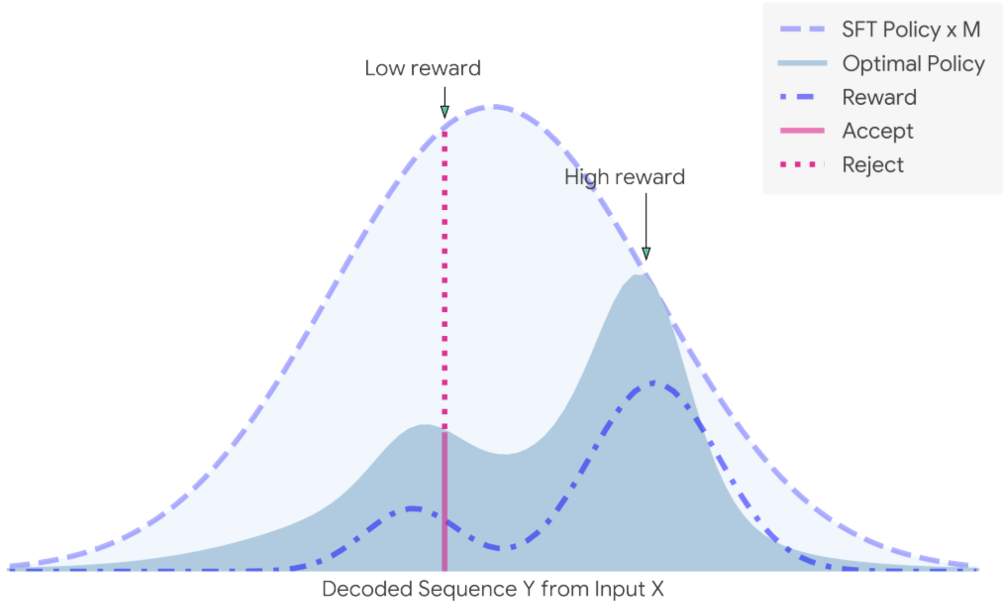
- Direct: directly fit the policy on D_{hf}.
- SFT-sample-rank: use $\pi_{sft}(y \mid x)$ to sample response pairs given prompts from the SFT training set and label them by a pre-trained reward model $r_\psi(x, y)$.
- RSO-sample-rank: use $\pi_{r_\psi}(y \mid x)$ induced by $r_\psi(x, y)$ to sample response pairs given prompts labelled by the pre-trained reward model $r_\psi(x, y)$, according to

$$\pi_r(y|x) = \frac{1}{Z(x)} \pi_{\text{sft}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)$$

Statistically speaking, "rso-sample-rank" is closer to $\pi^\star(y \mid x)$ than other two choices.

How to sample from $\pi_{r_\psi}(y \mid x)$? Rejection Sampling!

1. Start with empty $\mathcal{Y} = \{\}$.
2. Generate $y \sim \pi_{\text{sft}}(y|x)$ that is not in $\mathcal{Y}$ and $u \sim U[0, 1]$.
3. Let $M = \min\{m \mid m\pi_{\text{sft}}(y|x) \geq \pi_{r_\psi}(y|x) \text{ for all } y \notin \mathcal{Y}\}$[6]. If $u < \frac{\pi_{r_\psi}(y|x)}{M\pi_{\text{sft}}(y|x)}$, then we accept y and add it to $\mathcal{Y}$. Otherwise, we reject y.
4. Repeat step 2 and 3 until we get enough $\mathcal{Y}$.



Experiments

- Two tasks:
  - Reddit TL;DR summarization
  - AnthropicHH dialogue

Approach	Ablation		Metrics		
	Loss	Preference Pair	Proxy Reward (%)	Gold Reward (%)	AutoSxS (%)
Reddit TL;DR					
RAFT	cross-entropy	-	74.84	68.51	53.77
ReST	cross-entropy	-	49.03	46.17	34.36
DPO	sigmoid-norm	direct	84.35	76.09	67.72
	sigmoid-norm	sft-sample-rank	88.63	78.14	69.02
$RSO_{sigmoid-norm}$	sigmoid-norm	rso-sample-rank	**92.37**	**82.22**	**71.86**
AnthropicHH					
RAFT	cross-entropy	-	58.21	40.00	24.99
ReST	cross-entropy	-	43.48	30.33	15.58
DPO	sigmoid-norm	direct	51.63	36.13	24.01
	sigmoid-norm	sft-sample-rank	85.09	58.65	39.56
$RSO_{sigmoid-norm}$	sigmoid-norm	rso-sample-rank	**86.94**	**59.15**	**40.98**

Online Method:
Reinforced Self-Training (ReST)

Gulcehre, Caglar, et al. "Reinforced self-training (rest) for language modeling." *arXiv preprint arXiv:2308.08998* (2023).

Overview

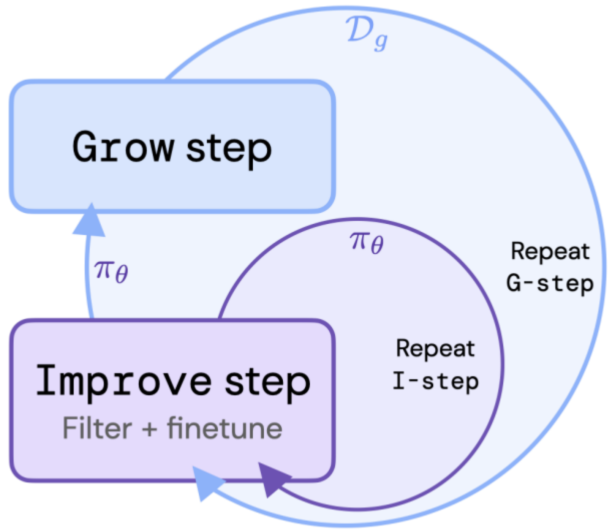


Figure 1 | ***ReST* method.** During `Grow` step, a policy generates a dataset. At `Improve` step, the filtered dataset is used to fine-tune the policy. Both steps are repeated, `Improve` step is repeated more frequently to amortise the dataset creation cost.

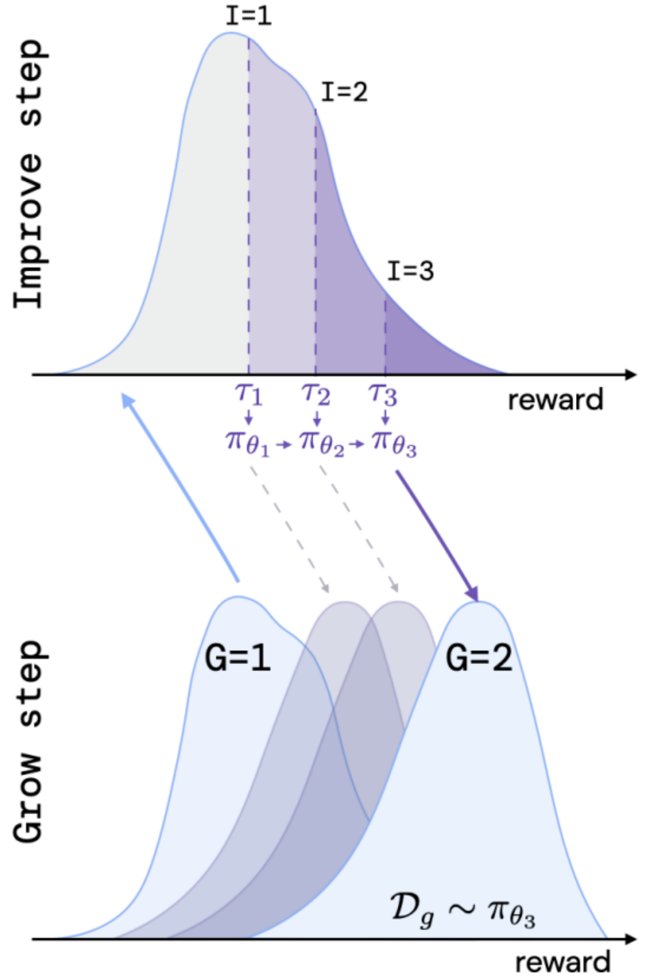


Figure 2 | ***ReST* algorithm.**

Reinforced Self-Training Algorithm

- Grow step (data generation): create an augmented dataset D_g by sampling many output sequences from the current policy π_θ

  i.e. $\boldsymbol{y} \sim \pi_\theta(\boldsymbol{y}|\boldsymbol{x})$ for $\boldsymbol{x} \sim \mathcal{D}$

  Then score the new dataset with a reward function $R(x, y)$.

- Improve step (policy improvement): use the dataset D_g to fine-tune the policy π_θ.

  Define a filtering function that t includes only samples with rewards higher than a certain threshold

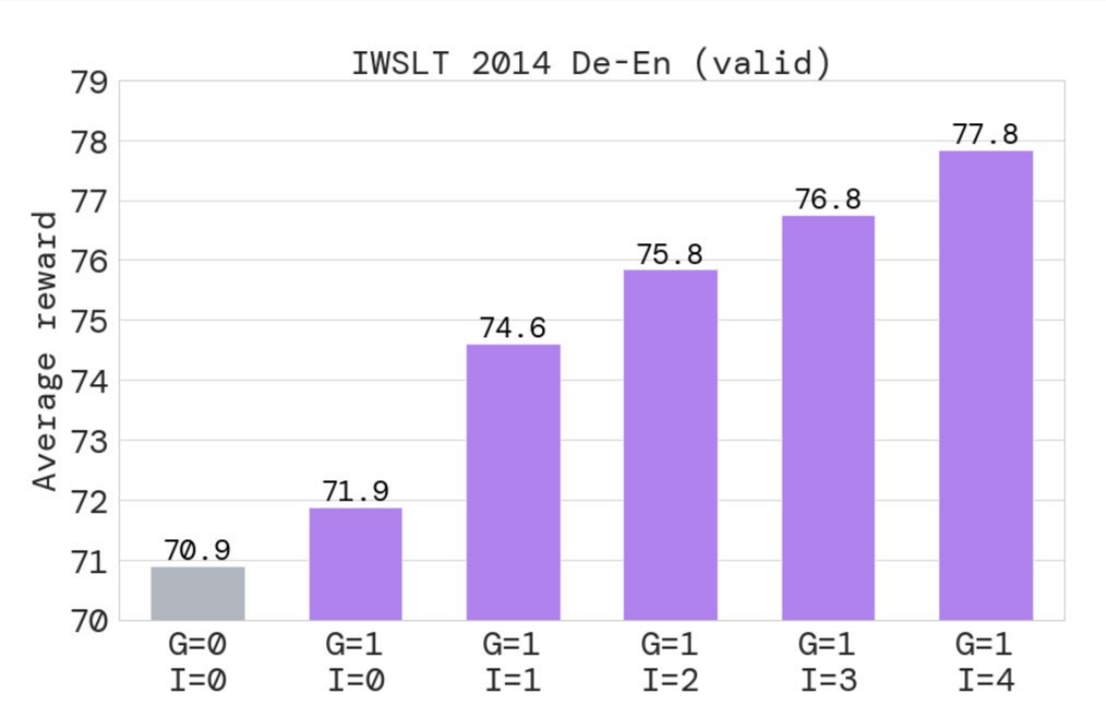
$$F(\boldsymbol{x}, \boldsymbol{y}; \tau) = \mathbb{1}_{R(\boldsymbol{x},\boldsymbol{y})>\tau}.$$

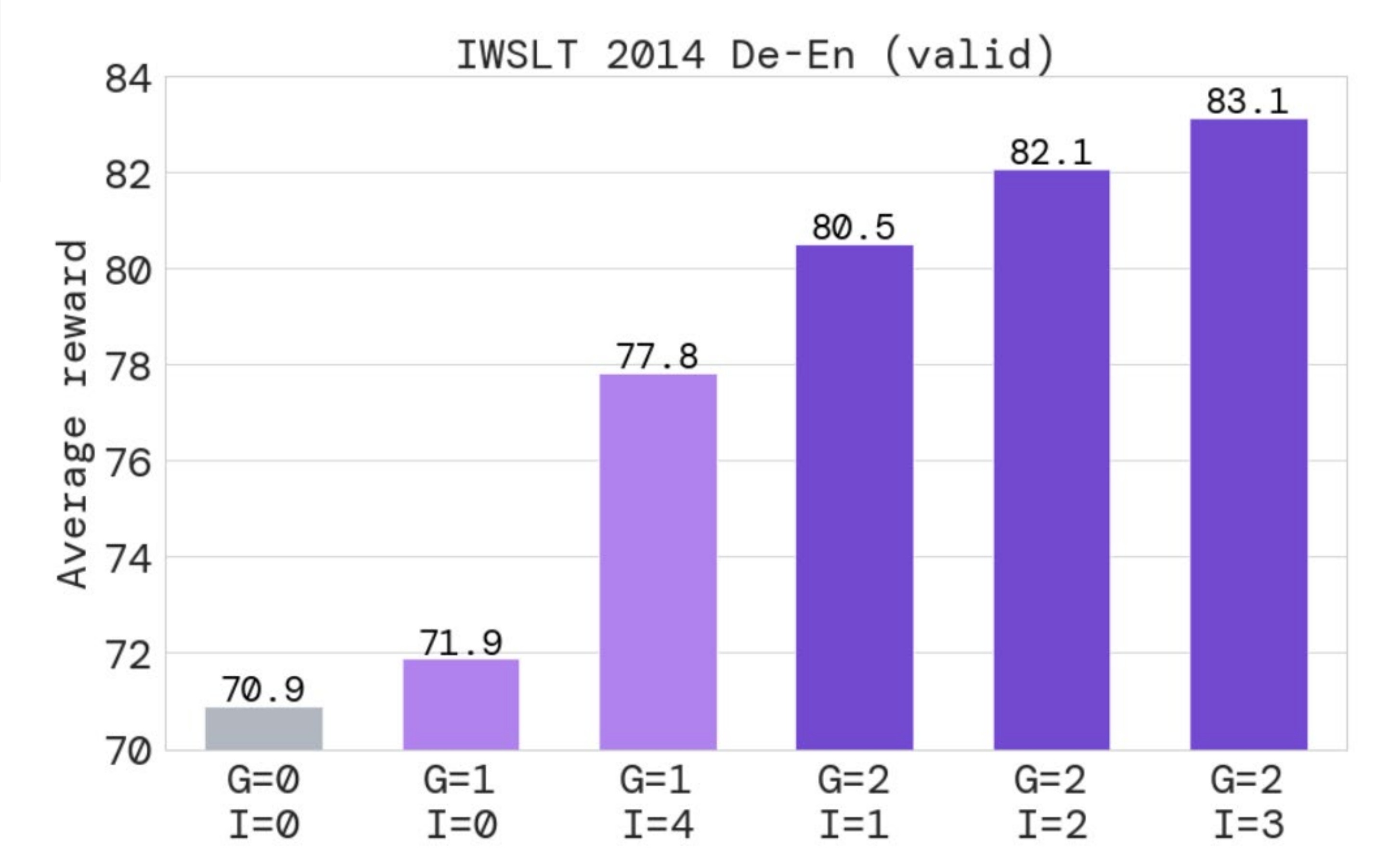
  Then finetune the current policy with an offline RL loss on the filtered data

$$J(\theta) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_g}\left[F(\boldsymbol{x}, \boldsymbol{y}; \tau)\ \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}; \theta)\right].$$

Experiments

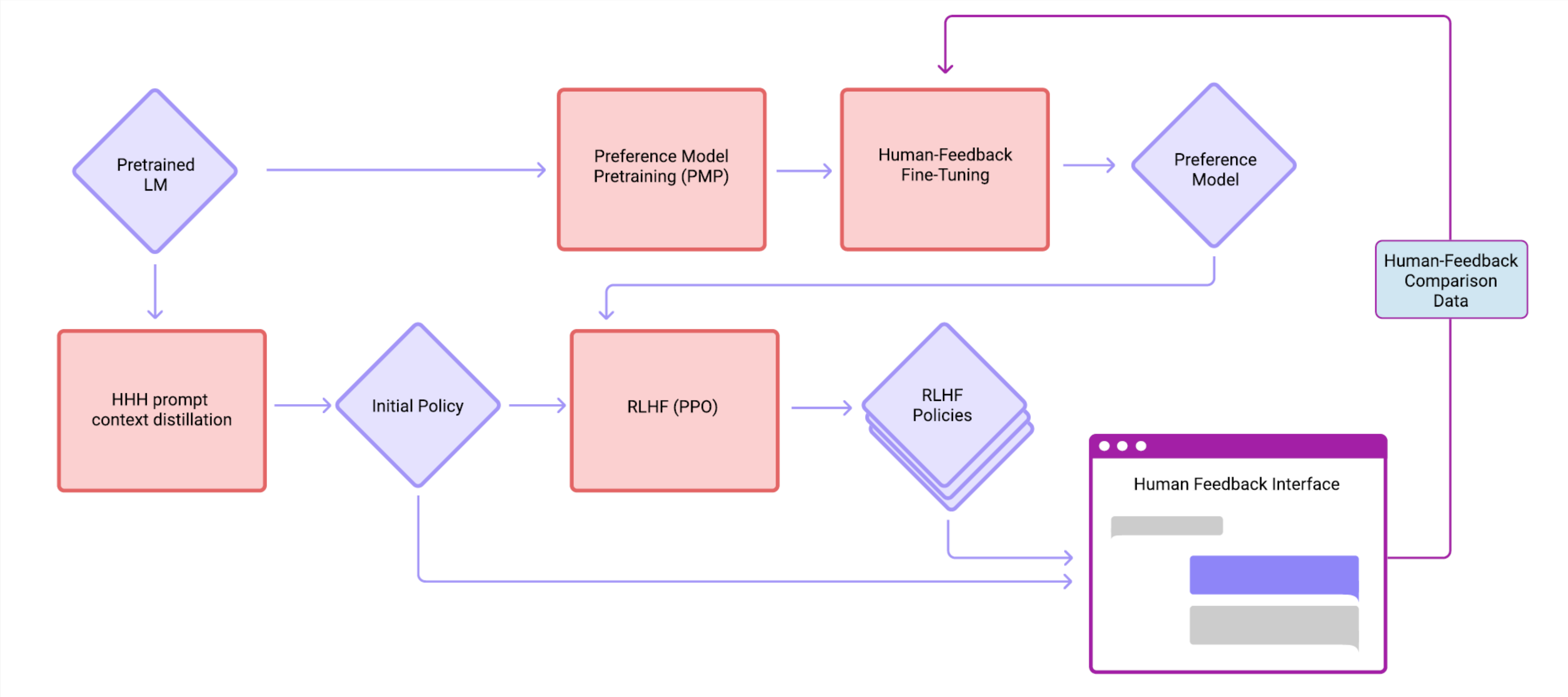
- Task: Translation





Discussion

Non-static Environment



Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[J]. arXiv preprint arXiv:2204.05862, 2022.

Future Direction

- 给定人工标注的budget，如何设计整个训练过程?