Direct Preference Optimization

2024.1.12

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov[*†] **Archit Sharma**[*†] **Eric Mitchell**[*†]

Stefano Ermon[†‡] **Christopher D. Manning**[†] **Chelsea Finn**[†]

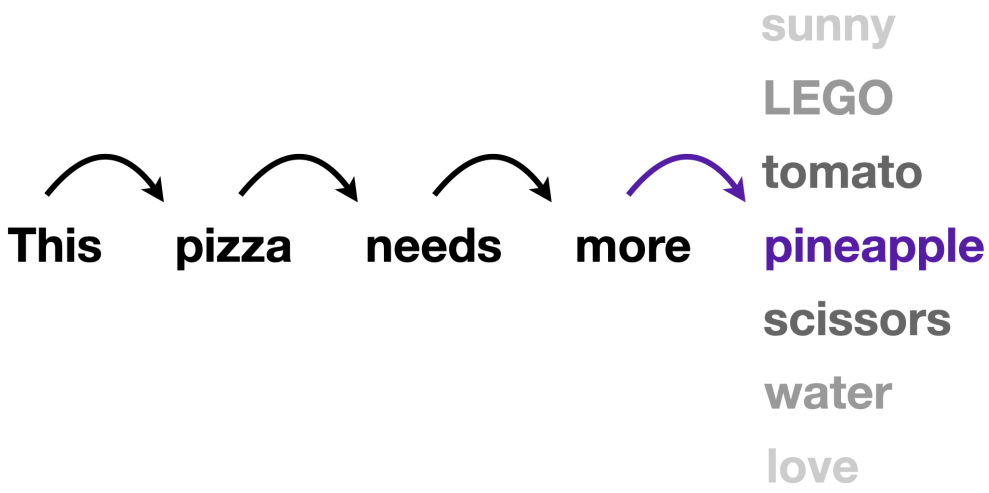
[†]Stanford University [‡]CZ Biohub
{rafailov,architsh,eric.mitchell}@cs.stanford.edu

Conference={NIPS2023},
Arxiv={2305.18290},
Citations={136}

Revision

LLM Training

1. Pretrain
2. SFT (Supervised Fine-tuning)
3. RLHF (Reinforcement Learning from Human Feed Back)



样例

```
{
  "instruction": "给定一个文字输入，将其中的所有数字加1。\n“明天的会议在9点开始，记得准时到达。
  "input": "",
  "output": "“明天的会议在10点开始，记得准时到达。”"
}
```

字段：

```
instruction: 指令
input: 输入（本数据集均为空）
output: 输出
```

Prompt

Explain the moon landing to a 6 year old in a few sentences.

用几句话给一个6岁的小孩解释一下“登月”

Completion

GPT-3

```
Explain the theory of gravity to a 6 year old.
```

给一个6岁小孩解释一下“重力”

```
Explain the theory of relativity to a 6 year old in a few sentences.
```

用几句话给一个6岁小孩解释一下“相对论”

```
Explain the big bang theory to a 6 year old.
```

给一个6岁小孩解释一下“宇宙大爆炸”

```
Explain evolution to a 6 year old.
```

给一个6岁小孩解释一下“进化”

InstructGPT

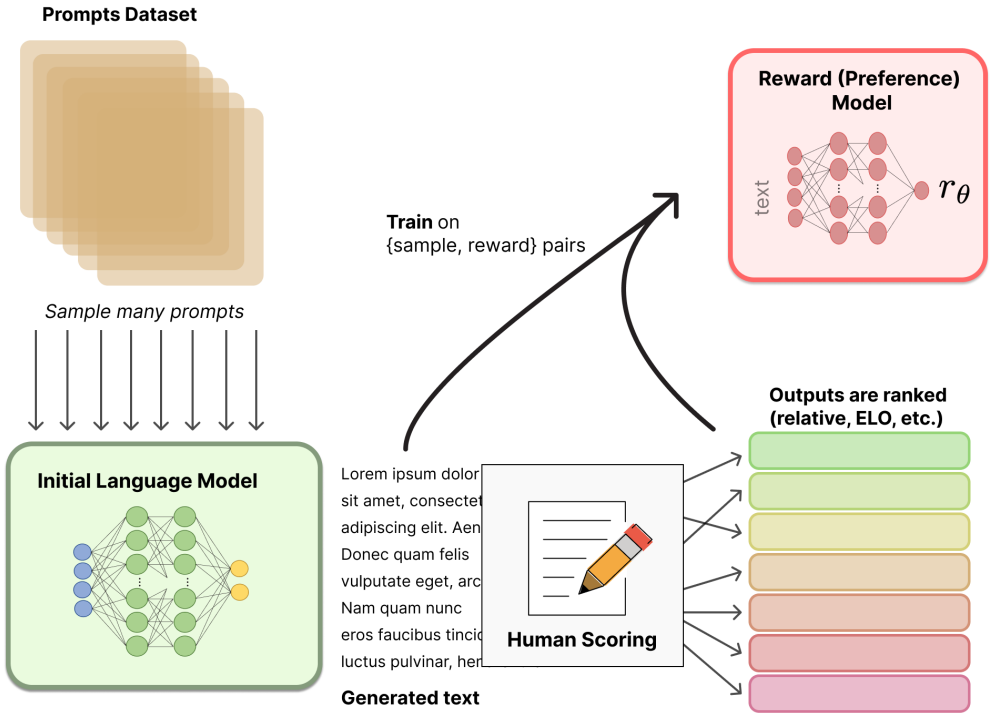
人们到月球上去拍照，并将拍好的照片带回地球。

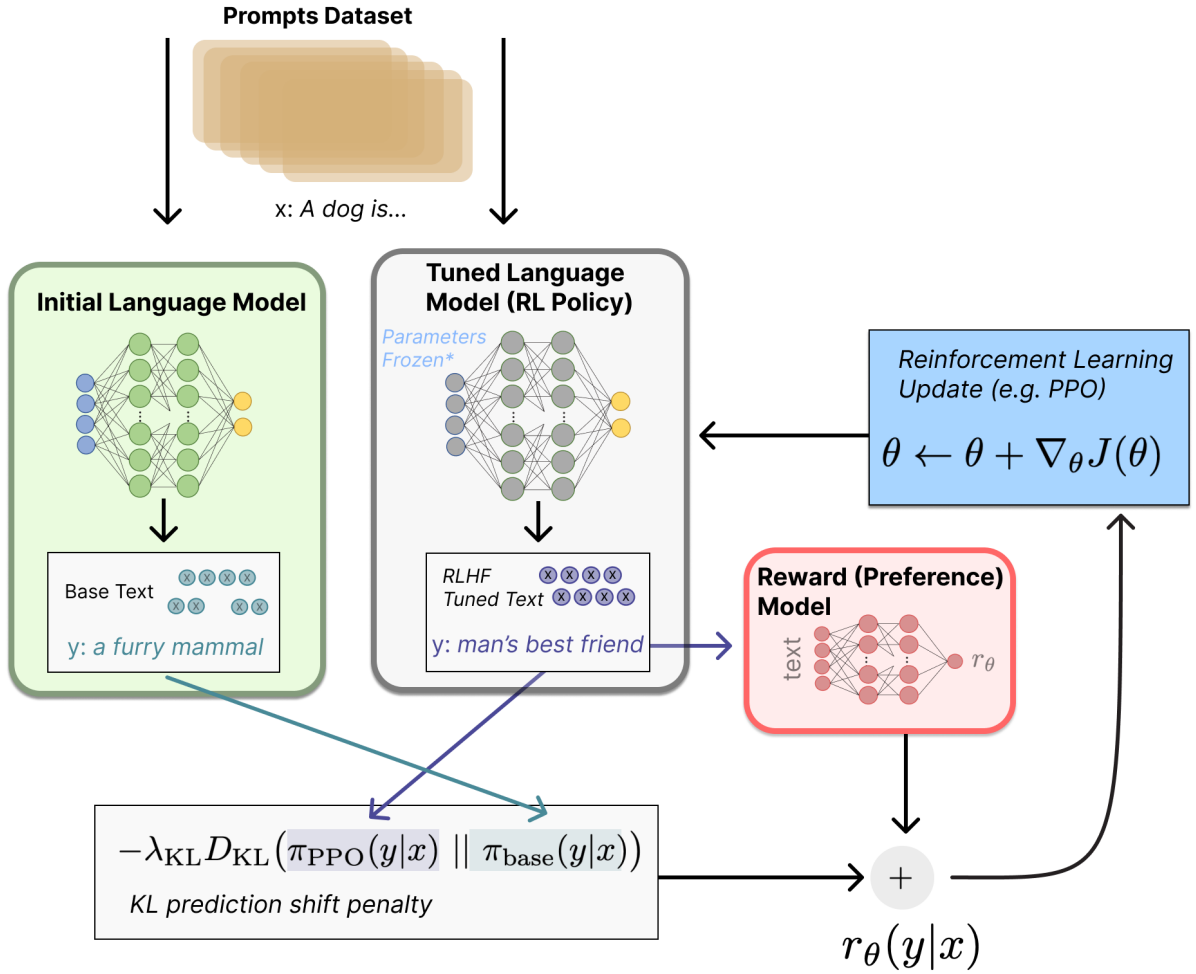
```
People went to the moon, and they took pictures of what they saw, and
sent them back to the earth so we could all see them.
```

RLHF

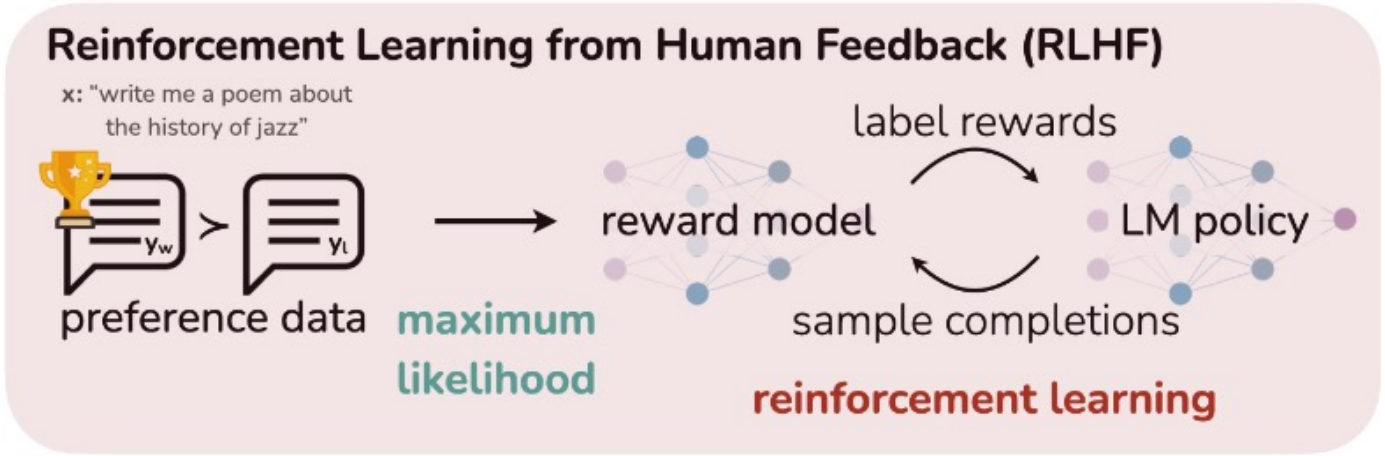
1. Generate Answers
2. Human Scoring
3. Training Reward Model
4. Reinforcement Learning (e.g. PPO)

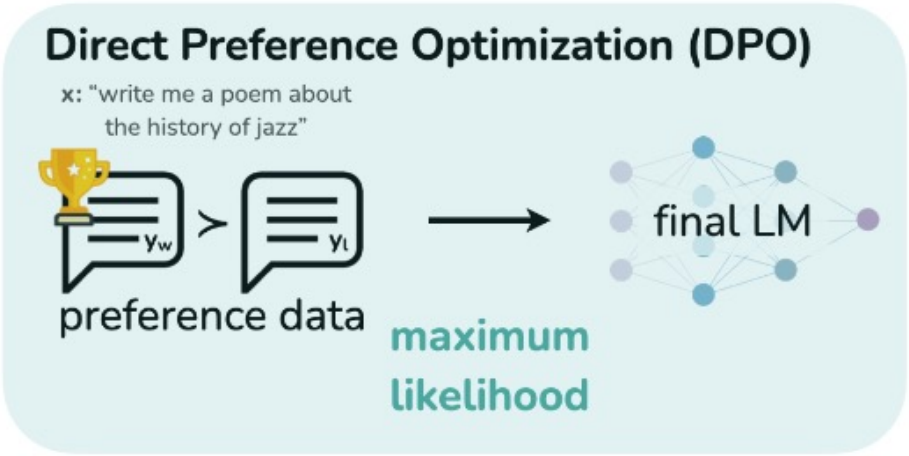
$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \,||\, \pi_{\mathrm{ref}}(y \mid x)\right]$$





Pipeline





Question:
"hello",
"how are you",
"What is your name?",
"What is your name?",
"Which is the best programming language?",
"Which is the best programming language?",
"Which is the best programming language?",

Chosen:
"hi nice to meet you",
"I am fine",
"My name is Mary",
"My name is Mary",
"Python",
"Python",
"Java",

Reject:
"leave me alone",
"I am not fine",
"Whats it to you?",
"I dont have a name",
"Javascript",
"C++",
"C++",

Preference Model

Bradley-Terry Model

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

Plackett-Luce Model

$$p^*(\tau|y_1, \ldots, y_K, x) = \prod_{k=1}^{K} \frac{\exp(r^*(x, y_{\tau(k)}))}{\sum_{j=k}^{K} \exp(r^*(x, y_{\tau(j)}))}$$

Reward Model Loss Function

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

Optimal Solution

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

$$= \max_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[r(x,y) - \beta\log\frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)} - \frac{1}{\beta}r(x,y)\right]$$

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

Define

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

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

$\pi^*(y|x) \geq 0$ for all y and $\sum_y \pi^*(y|x) = 1$.

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

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

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

Loss Function

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

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

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

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

Loss Function

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

$$\nabla_\theta\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$

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

DPO outline. The general DPO pipeline is as follows: 1) Sample completions $y_1, y_2 \sim \pi_{\text{ref}}(\cdot \mid x)$ for every prompt x, label with human preferences to construct the offline dataset of preferences $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l)^{(i)}\}_{i=1}^N$ and 2) optimize the language model π_θ to minimize $\mathcal{L}_{\text{DPO}}$ for the given π_{ref} and $\mathcal{D}$ and desired β. In practice, one would like to reuse preference datasets publicly available, rather than generating samples and gathering human preferences. Since the preference datasets

Theoretical Analysis

Definition 1. *We say that two reward functions* $r(x, y)$ *and* $r'(x, y)$ *are equivalent iff* $r(x, y) - r'(x, y) = f(x)$ *for some function* f.

It is easy to see that this is indeed an equivalence relation, which partitions the set of reward functions into classes. We can state the following two lemmas:

Lemma 1. *Under the Plackett-Luce, and in particular the Bradley-Terry, preference framework, two reward functions from the same class induce the same preference distribution.*

Lemma 2. *Two reward functions from the same equivalence class induce the same optimal policy under the constrained RL problem.*

$$
\begin{aligned}
p_{r'}(\tau|y_1, \ldots, y_K, x) &= \prod_{k=1}^{K} \frac{\exp(r'(x, y_{\tau(k)}))}{\sum_{j=k}^{K} \exp(r'(x, y_{\tau(j)}))} \\
&= \prod_{k=1}^{K} \frac{\exp(r(x, y_{\tau(k)}) + f(x))}{\sum_{j=k}^{K} \exp(r(x, y_{\tau(j)}) + f(x))} \\
&= \prod_{k=1}^{K} \frac{\exp(f(x)) \exp(r(x, y_{\tau(k)}))}{\exp(f(x)) \sum_{j=k}^{K} \exp(r(x, y_{\tau(j)}))} \\
&= \prod_{k=1}^{K} \frac{\exp(r(x, y_{\tau(k)}))}{\sum_{j=k}^{K} \exp(r(x, y_{\tau(j)}))} \\
&= p_r(\tau|y_1, \ldots, y_K, x),
\end{aligned}
$$

$$
\begin{aligned}
\pi_{r'}(y|x) &= \frac{1}{\sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r'(x, y)\right)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r'(x, y)\right) \\
&= \frac{1}{\sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta}(r(x, y) + f(x))\right)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta}(r(x, y) + f(x))\right) \\
&= \frac{1}{\exp\left(\frac{1}{\beta} f(x)\right) \sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right) \exp\left(\frac{1}{\beta} f(x)\right) \\
&= \frac{1}{\sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right) \\
&= \pi_r(y|x),
\end{aligned}
$$

Theoretical Analysis

Theorem 1. *Under mild assumptions, all reward classes consistent with the Plackett-Luce (and Bradley-Terry in particular) models can be represented with the reparameterization* $r(x, y) = \beta \log \frac{\pi(y|x)}{\pi_{ref}(y|x)}$ *for some model* $\pi(y \mid x)$ *and a given reference model* $\pi_{ref}(y \mid x)$.

We define the projection f as $f(r; \pi_{\text{ref}}, \beta)(x, y) = r(x, y) - \beta \log \sum_y \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x, y)\right)$

$$r'(x, y) = f(r, \pi_{\text{ref}}, \beta)(x, y) = r(x, y) - \beta \log Z(x) = \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)}$$

Proof. We will proceed using proof by contradiction. Assume we have two reward functions from the same class, such that $r'(x, y) = r(x, y) + f(x)$. Moreover, assume that $r'(x, y) = \beta \log \frac{\pi'(y|x)}{\pi_{\text{ref}}(y|x)}$ for some model $\pi'(y|x)$ and $r(x, y) = \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}$ for some model $\pi(y|x)$, such that $\pi \neq \pi'$. We then have

$$r'(x, y) = r(x, y) + f(x) = \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} + f(x) = \beta \log \frac{\pi(y|x) \exp(\frac{1}{\beta} f(x))}{\pi_{\text{ref}}(y|x)} = \beta \log \frac{\pi'(y|x)}{\pi_{\text{ref}}(y|x)}$$

for all prompts x and completions y. Then we must have $\pi(y|x) \exp(\frac{1}{\beta} f(x)) = \pi'(y|x)$. Since these are distributions, summing over y on both sides, we obtain that $\exp(\frac{1}{\beta} f(x)) = 1$ and since $\beta > 0$, we must have $f(x) = 0$ for all x. Therefore $r(x, y) = r'(x, y)$. This completes the proof. □

$$\sum_y \underbrace{\pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x, y)\right)}_{=\pi(y|x), \text{ using Thm. 1 reparam.}} = 1$$

Experiment

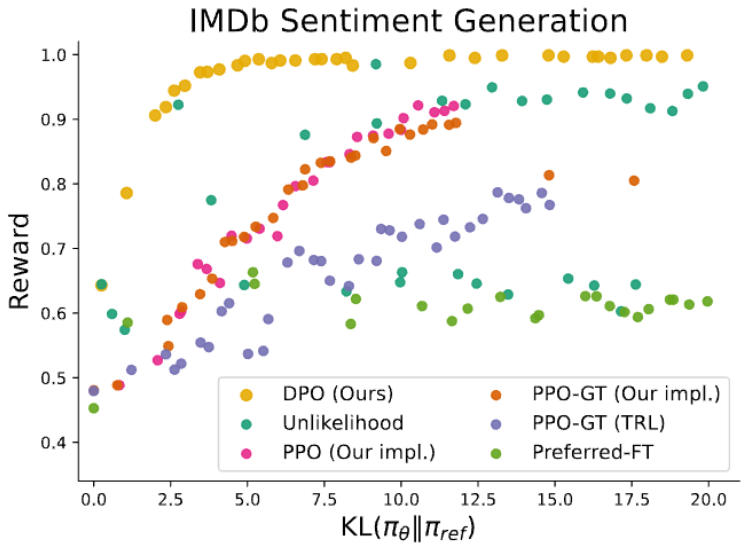
Task:

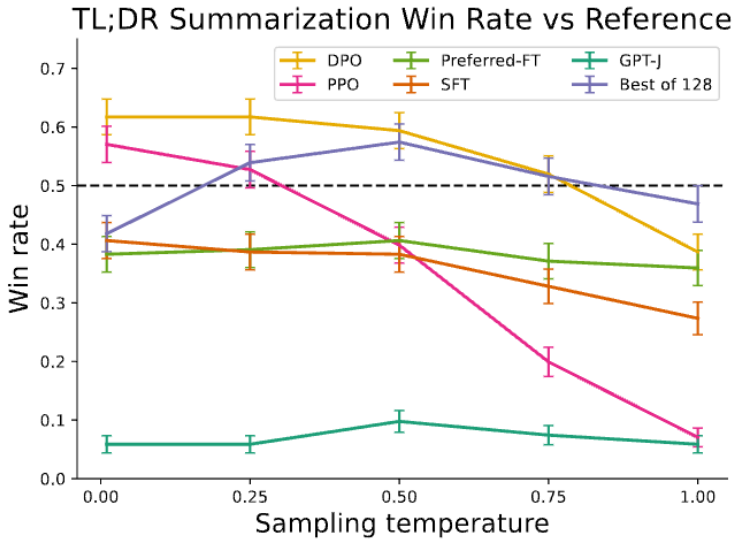
1. Controlled Sentiment Generation
2. Summarization
3. Single-turn Dialogue

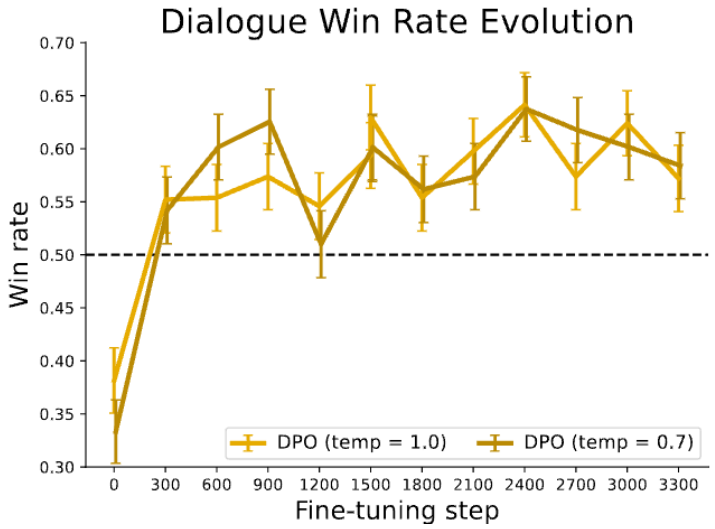
	Win rate vs. ground truth	
Alg.	Temp 0	Temp 0.25
DPO	0.36	0.31
PPO	0.26	0.23

Table 1: GPT-4 win rates vs. ground truth summaries for out-of-distribution CNN/DailyMail input articles.

	DPO	SFT	PPO-1
N respondents	272	122	199
GPT-4 (S) win %	47	27	13
GPT-4 (C) win %	54	32	12
Human win %	58	43	17
GPT-4 (S)-H agree	70	77	86
GPT-4 (C)-H agree	67	79	85
H-H agree	65	-	87







THANK YOU