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

Подготовил
Федоров Никита БПМИ202

nmfedorov@edu.hse.ru

Recap: LLMs training pipeline

1) unsupervised training
2) supervised fine-tuning (SFT) -> $\pi^{\text{SFT}}(y \mid x)$
3) reward model training
   a) sampling: $(y_1, y_2) \sim \pi^{\text{SFT}}(y \mid x)$
   b) human preferences markup: $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$
   c) train reward model $r_\phi(x, y)$ with loss $\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \boxed{\sigma(r_\phi(x, y_w) - r_\phi(x, y_l))}\right]$

reward model estimates probability
that y_w is better than y_l

4) RL (next slide)

Recap: Reinforcement Learning on Human Feedback (RLHF)

Optimization task:

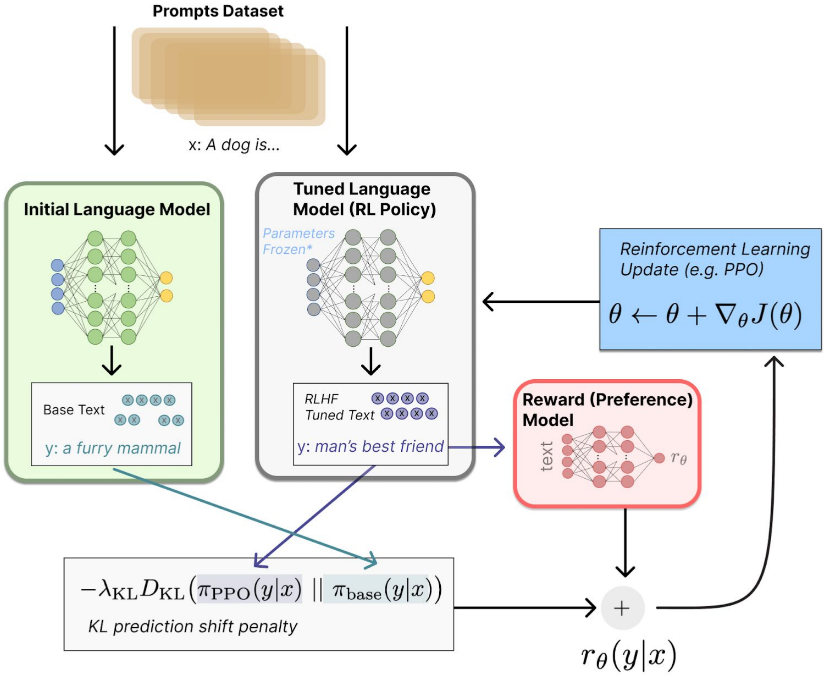
$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \mid\mid \pi_{\mathrm{ref}}(y \mid x)\right]$$

Not differentiable, so construct reward

$$r(x,y) = r_\phi(x,y) - \beta(\log \pi_\theta(y \mid x) - \log \pi_{\mathrm{ref}}(y \mid x))$$

and maximize it using PPO (proximal

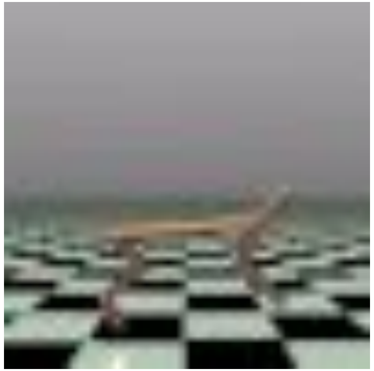
policy optimization - RL algorithm)

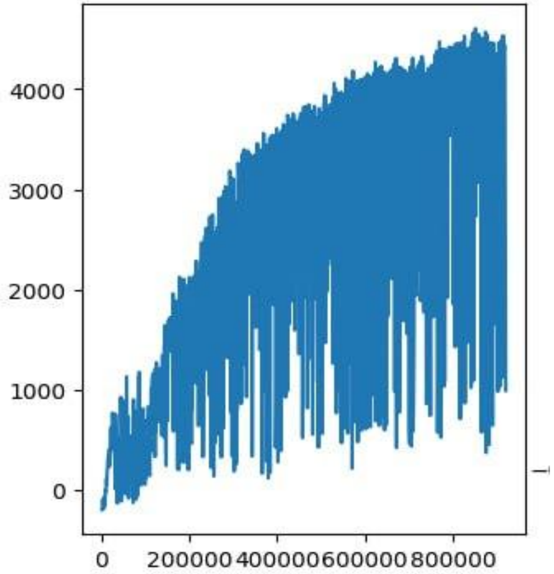


Problems of RLHF

expected PPO

real PPO





Direct Preference Optimization

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

The exact solution: $$\pi_r(y \mid x) = \frac{1}{Z(x)}\pi_{\mathrm{ref}}(y \mid x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$, where $Z(x) = \sum_y \pi_{\mathrm{ref}}(y \mid x)\exp\left(\frac{1}{\beta}r(x,y)\right)$

DPO

Express reward through optimal policy: $$r(x,y) = \beta\log\frac{\pi_r(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} + \beta\log Z(x)$$

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

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

LLM DPO pipeline

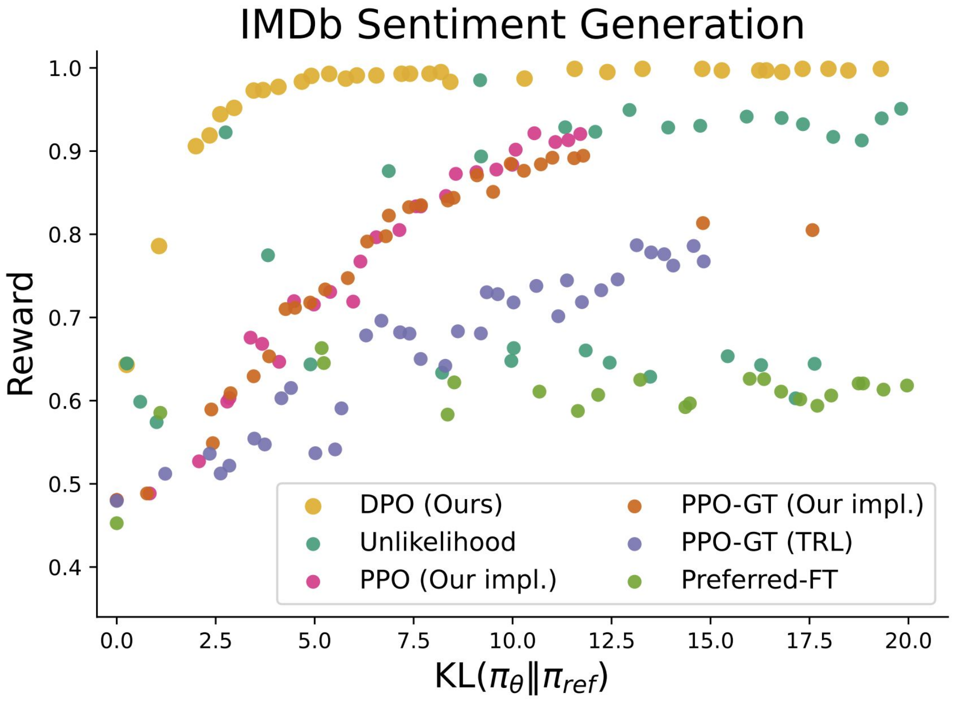
1) unsupervised training

2) supervised fine-tuning (SFT) -> $\pi^{\text{SFT}}(y \mid x)$

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

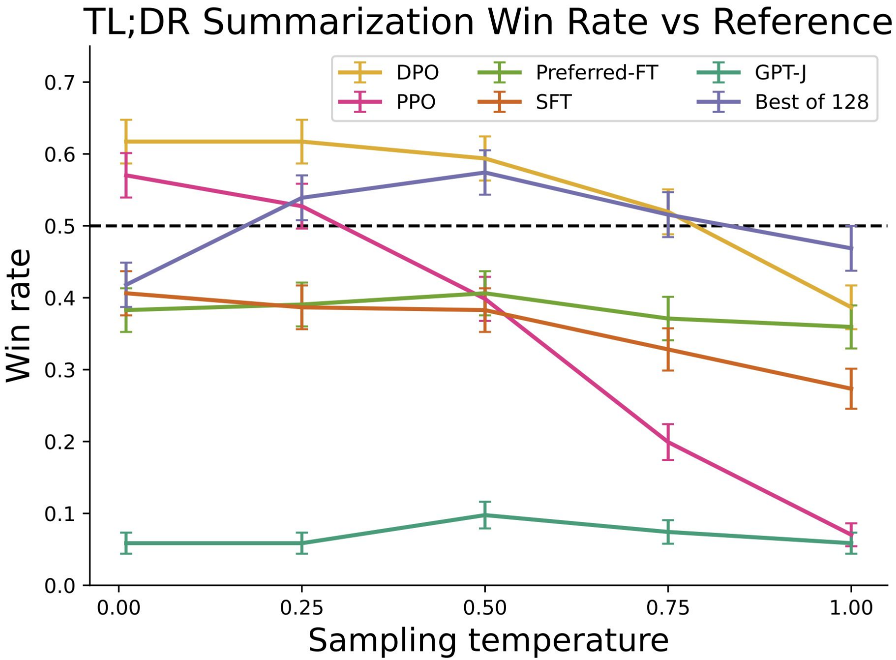
Experiments: Controlled sentiment generation task

- LM is tasked to continue prompts with positive sentiment review
- IMDB dataset
- preference dataset is marked up with pretrained sentiment classifier
- methods are compared by the Reward-KL frontier (see plot)



Experiments: summarization task

- LM is tasked to summarize Reddit's posts
- Reddit TL;DR summarization dataset
- methods are compared by **win rate** against reference summarization
- GPT-4 is used for win rate calculation



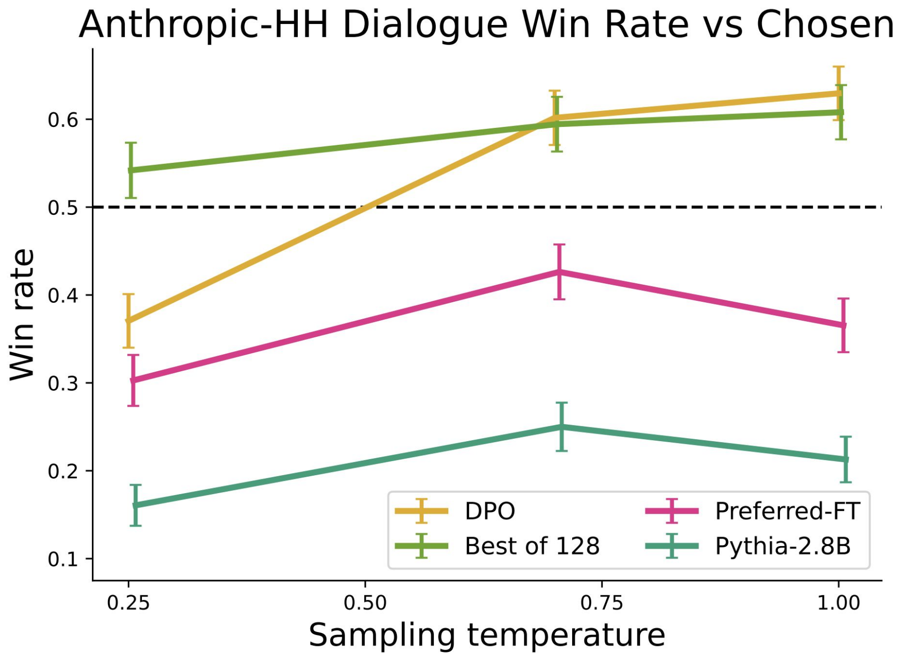
Experiments: summarization on OOD

	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.

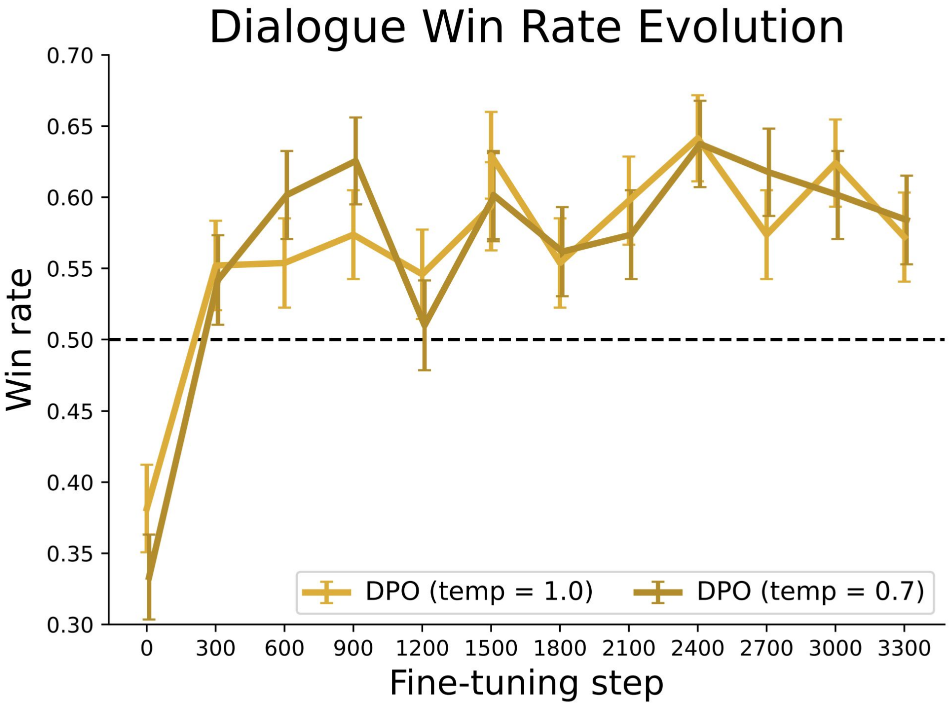
Experiments: single-turn dialogue task

- LM is tasked to give a response to human query
- Anthropic Helpful and Harmless dialogue dataset (Anthropic-HH)
- methods are compared by **win rate** against reference summarization
- GPT-4 is used for win rate calculation



Experiments: single-turn dialogue task

- experiment demonstrating stability of training with DPO



Источники

- статья: https://arxiv.org/pdf/2305.18290.pdf