大模型强化学习最新进展

傅延赫

2024年7月12日

动机

- **处理目标复杂、定义不明确或难以精准表述的任务。**
- 例子：损失函数无法定义一个输出是否是“有趣”的，但对人类来说，评判模型生成的笑话是否有趣却很简单。

- **微调方法无法告诉模型什么不好。**
- 微调只能传递给模型什么是好的输出结果，模型只会一味的照着label进行拟合，但我们无法告诉模型坏的输出结果什么样。

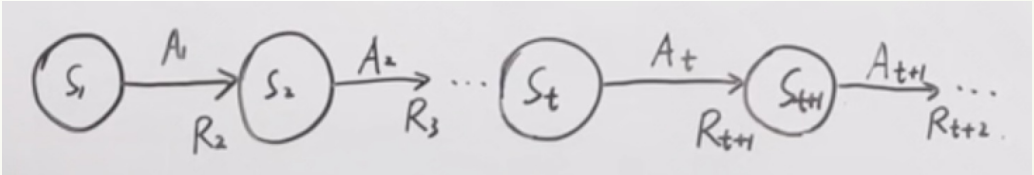
- **模型输出结果需要可控。**
- 小到模型重复生成、幻觉问题，大到歧视问题，我们需要一种手段控制模型输出。

强化学习基础

- 马尔可夫链：状态S
- 马尔可夫奖励过程：状态S、奖励R
- 马尔可夫决策过程：状态S、动作A、奖励R和环境的动态特性P

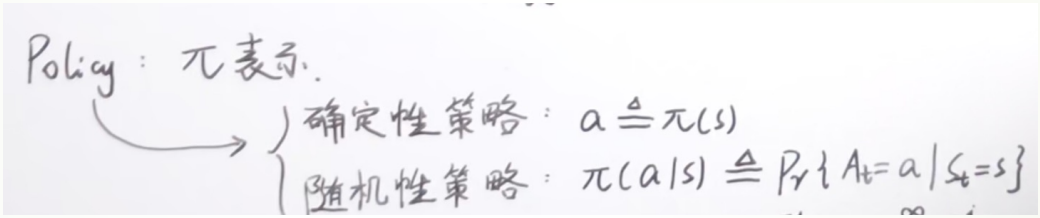
$$p(s', r \mid s, a) \triangleq Pr\{S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a\}$$

- 有向图模型：



强化学习基础

- 求解什么？
- 求解的是策略，输入某一时刻的状态，输出下一时刻的动作



- 目标是什么？
- 目标是该策略获得的奖励最大
- *注意：获得的奖励并不是某一个时刻的R，而是之后所有时刻的R，一般用回报G来表示

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-1} R_T = \sum_{i=0}^{\infty} \gamma^i R_{t+i+1}$$

$$\gamma \in [0,1]$$

强化学习基础

- 回报只是一条可能的轨迹的Reward和，真正的价值需要评估所有可能的轨迹

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-1} R_T = \sum_{i=0}^{\infty} \gamma^i R_{t+i+1}$$

$$\gamma \in [0, 1]$$

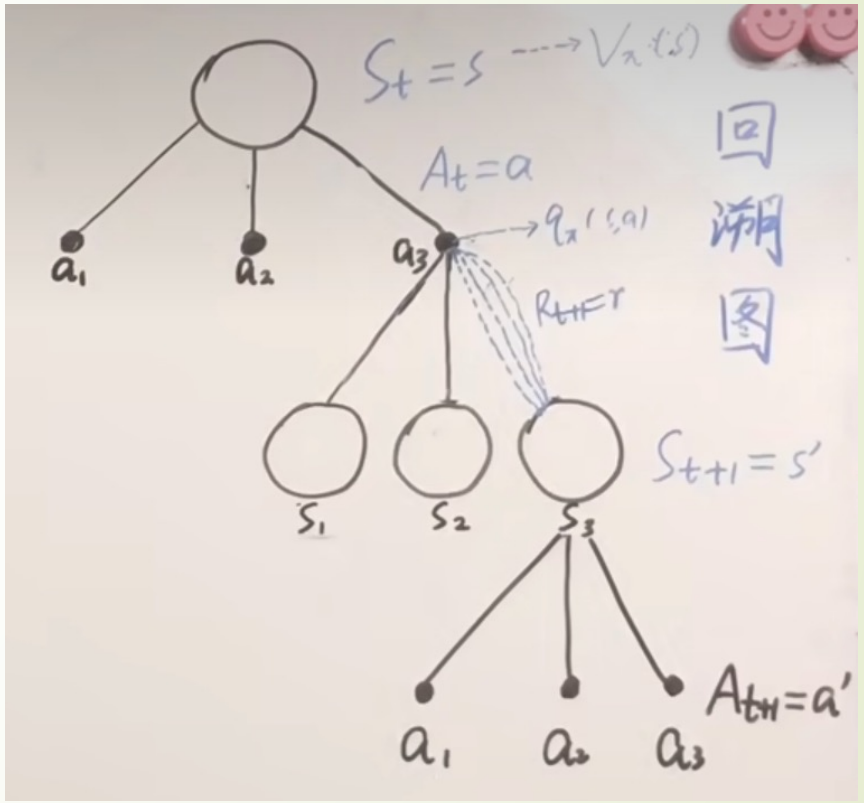
- 价值函数考虑了所有回报G的期望

$$\text{Value Function}: \begin{cases} v_\pi(s) \triangleq E_\pi[G_t \mid S_t = s] \\ q_\pi(s,a) \triangleq E_\pi[G_t \mid S_t = s, A_t = a] \end{cases}$$

- 扩展：状态价值函数和动作价值函数有显然的关系

$$\bullet\ v_\pi(s) = \sum_{a \in A} \pi(a|s) \cdot q_\pi(s,a)$$

$$\bullet\ q_\pi(s,a) = \sum_{r,s'} p(s', r \mid s, a) \cdot [r + \gamma\ v_\pi(s')]$$



LLM和强化学习

- 在自回归生成任务中：
  - 状态S就是每一步的完整文本，初始状态是输入的prompt
  - 动作A就是每次新生成哪个token
  - 奖励R就是人类视角下，生成的文本“好”还是“坏”

- 强化学习的难点在于如何预估回报G（奖励R）
- OpenAI：直接标数据寻一个可以输入文本、输出奖励值的模型

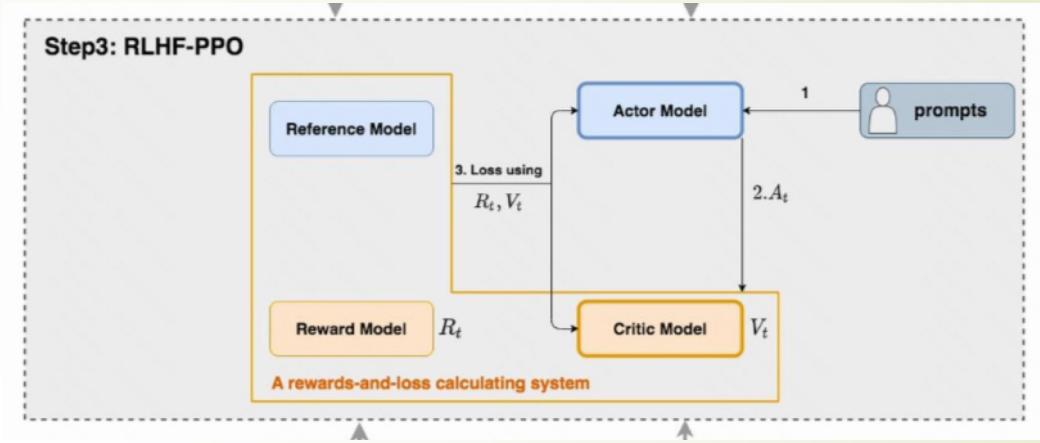
PPO

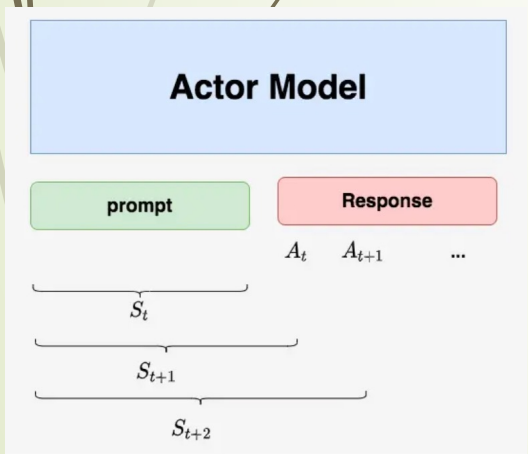
Actor Model：训练的目标语言模型，初始化为微调后的参数

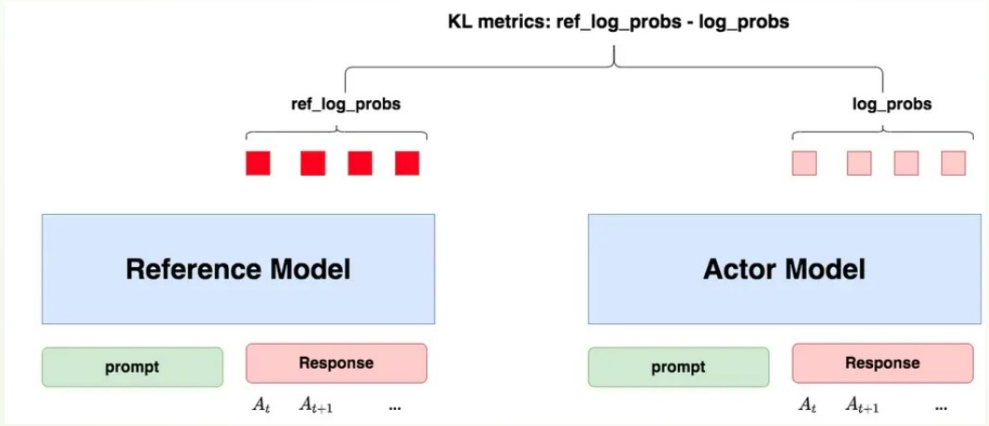
Critic Model：**预估价值** V $V_t = R_t + \gamma V_{t+1}$

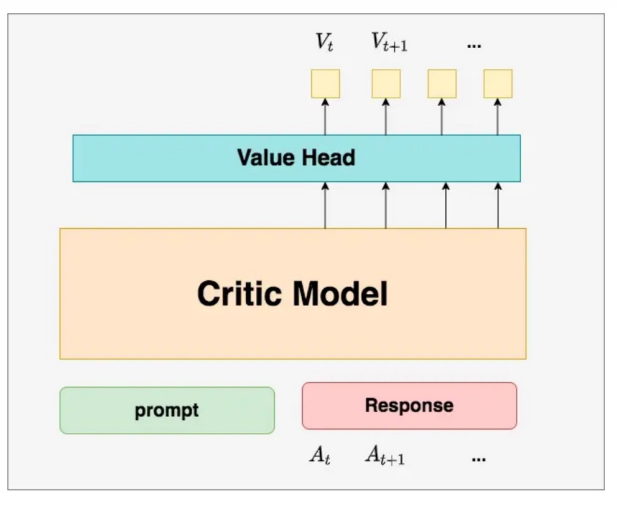
Reward Model*：**计算**即时奖励 R

Reference Model*：给语言模型增加一些“约束”，防止语言模型训歪









PPO

Actor的损失：使用优势取代价值，表示超出预估的价值

$$Adv_t = R_t + \gamma * V_{t+1} - V_t$$

将KL散度加到奖励R中：

$$\begin{cases} R_t = -kl\_ctl * (log\frac{P(A_t|S_t)}{P_{ref}(A_t|S_t)}), t \neq T \\ R_t = -kl\_ctl * (log\frac{P(A_t|S_t)}{P_{ref}(A_t|S_t)}) + R_t, t = T \end{cases}$$

当$t \neq T$时，我们更加关心Actor是否有在Ref的约束下生产token

当$t=T$时，我们不仅关心Actor是否遵从了Ref的约束，也关心真正的即时奖励R

为什么只有最后一个时刻的 Rt 被纳入了考量呢？这是因为在Reward模型训练阶段，就是用这个位置的 Rt 来表示对完整的prompt + response的奖励预测，然后用这个指标来做模型eval的。所以到了RLHF的场景下，其余时刻的即时奖励，我们就用“Actor是否遵循了Ref的约束”来进行评价。

$$actor\_loss = -Adv_t logP(A_t|S_t)$$

PPO

Critic的损失：

$$Critic\_loss = (R_t + \gamma * V_{t+1} - V_t)^2$$

包括：

1、Critic对t时刻的总收益的预估

2、Reward计算出的即时收益， Critic预测出的 t+1 及之后时候的收益的折现，这是比 Vt 更接近t时刻真值总收益的一个值

response一次完整生成完毕后，Actor和Critic进行一次参数更新

DPO

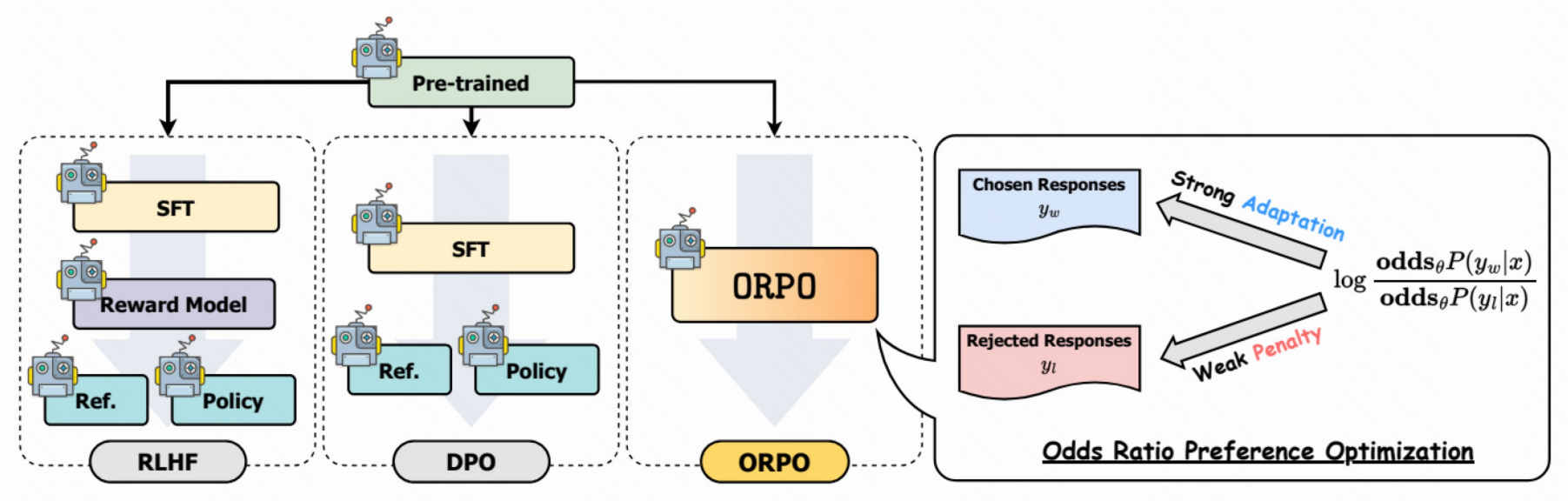
- PPO耗时：需要提前训练好RM
- PPO占内存：需要加载4个模型
- PPO标数据很麻烦：需要标每条的绝对价值
- DPO只有两个模型（reward省了），SFT后监督训练即可
- 损失函数：

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

- 数据构造：
  - {“prompt”: “xxx”, “chosen”: “xxx”, “reject”: “xxx”}
  - 通过采样SFT模型并人工标注pair对得到

ORPO: Monolithic Preference Optimization without Reference Model

- PPO和DPO要两步骤完成偏好对齐，并且需要初始化一个参考模型到GPU中



- 本方法设计了一个损失函数加入到SFT损失中，将对齐工作与微调同时进行，节约时间与显存成本

ORPO: Monolithic Preference Optimization without Reference Model

- 交叉熵损失函数只能增大标准答案的概率，而对被拒绝的句子没有惩罚

$$\mathcal{L} = -\frac{1}{m}\sum_{k=1}^{m}\log P(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$$

$$= -\frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{|V|} y_i^{(k)} \cdot \log(p_i^{(k)})$$

- 设计损失惩罚被拒绝的句子：

$$\mathbf{odds}_\theta(y|x) = \frac{P_\theta(y|x)}{1 - P_\theta(y|x)}$$

$$\mathcal{L}_{OR} = -\log\sigma\left(\log\frac{\mathbf{odds}_\theta(y_w|x)}{\mathbf{odds}_\theta(y_l|x)}\right)$$

$$\mathbf{OR}_\theta(y_w, y_l) = \frac{\mathbf{odds}_\theta(y_w|x)}{\mathbf{odds}_\theta(y_l|x)}$$

$$\mathcal{L}_{ORPO} = \mathbb{E}_{(x,y_w,y_l)}\left[\mathcal{L}_{SFT} + \lambda \cdot \mathcal{L}_{OR}\right]$$

其中，OR表示相同输入下，生成chosen的概率是生成rejected的概率的几倍

ORPO: Monolithic Preference Optimization without Reference Model

- 实验：


```
{
  "instruction":"How did US states get their names?",
  "output":"US states got their names from a variety of sources includin
  "generator":"example",
  "dataset":"helpful_base",
  "datasplit":"eval"
},
{
  "instruction":"Hi, my sister and her girlfriends want me to play kickb
  "output":"Certainly! Kickball is similar to baseball, but instead of u
  "generator":"example",
  "dataset":"helpful_base",
  "datasplit":"eval"
},
```


Model Name	Size	$\text{AlpacaEval}_{1.0}$	$\text{AlpacaEval}_{2.0}$
Phi-2 + SFT	2.7B	48.37% (1.77)	0.11% (0.06)
Phi-2 + SFT + DPO	2.7B	50.63% (1.77)	0.78% (0.22)
Phi-2 + ORPO *(Ours)*	2.7B	**71.80% (1.59)**	**6.35% (0.74)**
Llama-2 Chat *	7B	71.34% (1.59)	4.96% (0.67)
Llama-2 Chat *	13B	81.09% (1.38)	7.70% (0.83)
Llama-2 + ORPO *(Ours)*	7B	**81.26% (1.37)**	**9.44% (0.85)**
Zephyr (α) *	7B	85.76% (1.23)	8.35% (0.87)
Zephyr (β) *	7B	90.60% (1.03)	10.99% (0.96)
Mistral-ORPO-α *(Ours)*	7B	87.92% (1.14)	11.33% (0.97)
Mistral-ORPO-β *(Ours)*	7B	**91.41% (1.15)**	**12.20% (0.98)**

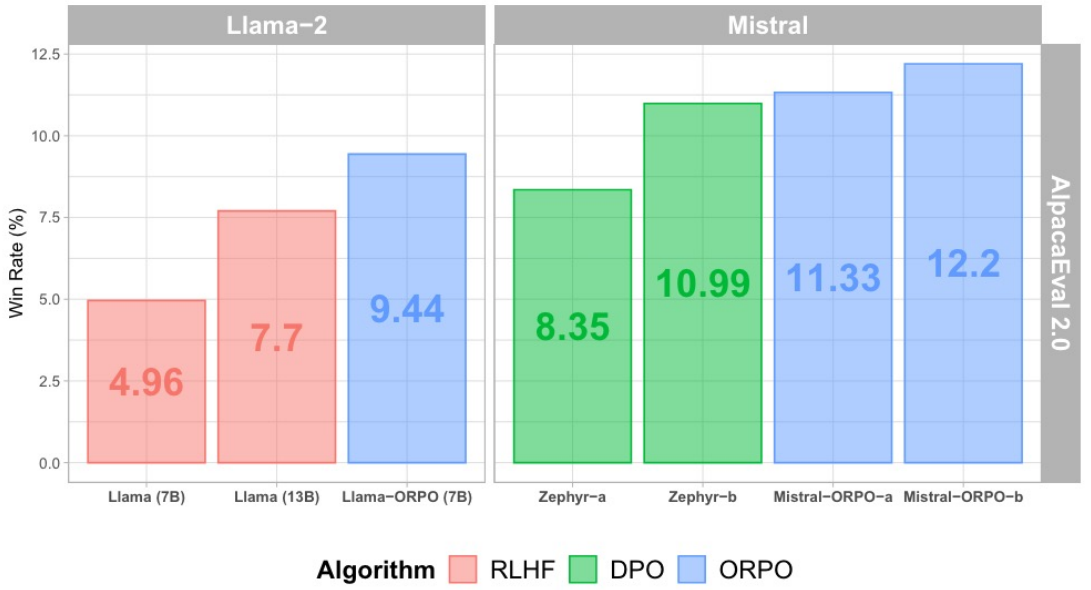


Figure 1: $\text{AlpacaEval}_{2.0}$ result of Llama-2 (7B) and Mistral (7B) fine-tuned with ORPO (blue) in comparison to the state-of-the-art models. Notably, Mistral-ORPO-α

KTO: Model Alignment as Prospect Theoretic Optimization

- **前景理论：为什么人类在面对不确定事件时会做出无法最大化期望值的决策？**
- 由于人类是厌恶损失的，假设一场赌博以 80% 的概率返回 100 美元，以 20% 的概率返回 0 美元，不参加赌博直接返回60美元。在数学上，避免赌博的收益期望为60美元；而接受赌博的收益期望为0.8 * 100+0.2 * 0=80美元。
- 大模型理论上应该选择期望最大的决策，即参加赌博；而大多数的人面对同样的场景则往往会选择不参加赌博直接获取60美元。
- 因为人有很多类似的无法做出期望最大决策的场景，所以用人标注的数据对 指导大模型的训练往往也会使大模型学到这种偏好，而无法做出期望收益更大的决策。
- 例子：“一个人下班之后很累，他期望用什么交通工具回家？”，标注者认为打车比坐公交好，所以模型学到打车比坐公交更适合回家。这是我们不想模型学到的一种错误的个体偏好。

KTO: Model Alignment as Prospect Theoretic Optimization

- 解决：拆解偏好数据
- 将原始的一条数据：(x, y_w, y_l) 变为两条数据：$(x, y_w, 1)$ 和 $(x, y_l, 0)$
- 除了消除错误偏好外，这样做还可以：
  - 节约了很大的数据收集成本
  - 在数据量少的情况下，也可以训练，不至于无法构造数据对
  - 在正负样本数据不平衡的情况下，也可以鲁棒
  - 如果基座够强大，KTO甚至可以不进行SFT

- 如何训练？

KTO: Model Alignment as Prospect Theoretic Optimization

- 前景理论模型
- 参考点：人们通常不是基于绝对结果做决策，而是基于相对于某个参考点的变化。参考点可以是当前的财富水平、预期的结果或其他某个基准。
- 价值函数：人们对收益和损失的感知和评价

$$v(z;\lambda,\alpha,z_0) = \begin{cases} (z-z_0)^\alpha & \text{if } z \geq z_0 \\ -\lambda(z_0-z)^\alpha & \text{if } z < z_0 \end{cases}$$

$$\alpha = 0.88 \qquad \lambda = 2.25$$

- 价值函数要求厌恶风险、边际递减

KTO: Model Alignment as Prospect Theoretic Optimization

- KTO

$$r_\theta(x,y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

$$z_0 = \mathbb{E}_{x' \sim D}\left[\mathrm{KL}(\pi_\theta(y'|x') \| \pi_{\text{ref}}(y'|x'))\right]$$

$$v(x,y) = \begin{cases} \lambda_D \sigma(\beta(r_\theta(x,y) - z_0)) \text{ if } y \sim y_{\text{desirable}}|x \\ \lambda_U \sigma(\beta(z_0 - r_\theta(x,y))) \text{ if } y \sim y_{\text{undesirable}}|x \end{cases}$$

- 每个样本的隐式reward和DPO一致
- 参考点假设人类是根据看过的所有(x,y)数据后来判断某个(x,y)的相对质量的，而不是其绝对质量。因此本文将参考点设为期望reward

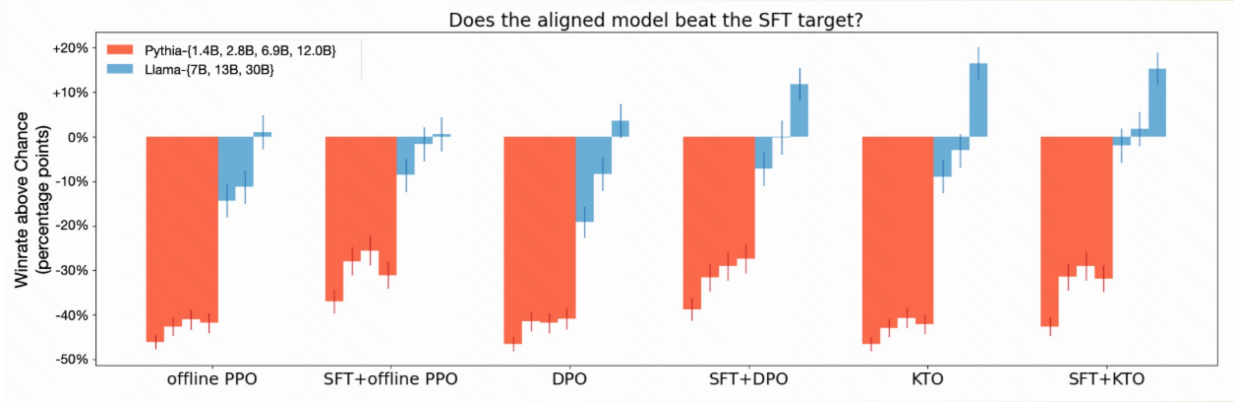
$$L_{\text{KTO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x,y \sim D}[\lambda_y - v(x,y)]$$

- 由于KTO是分别针对单条数据，如果数据是正样本，那么一定要超过参考点才会反馈预测正确；对于负样本，需要低于参考点才会反馈预测正确

KTO: Model Alignment as Prospect Theoretic Optimization

实验

Dataset (→) Metric (→)	MMLU EM	GSM8k EM	HumanEval pass@1	BBH EM
SFT	57.2	39.0	30.1	46.3
DPO	58.2	40.0	30.1	44.1
ORPO ($\lambda = 0.1$)	57.1	36.5	29.5	47.5
KTO ($\beta = 0.1, \lambda_D = 1$)	**58.6**	**53.5**	**30.9**	**52.6**
KTO (one-y-per-x)	58.0	50.0	30.7	49.9



ABCD选项问题+代码生成

Method	Winrate vs. SFT Target
Mistral-7B (unaligned)	0.525 ± 0.037
Mistral-7B + DPO	0.600 ± 0.037
Mistral-7B + KTO (all y per x)	**0.652 ± 0.036**
Mistral-7B + KTO (one y per x)	0.631 ± 0.036
Mistral-7B-Instruct	0.621 ± 0.031

1. KTO在1B到30B的模型中的性能表现超越了PPO和DPO。
2. KTO可以在减少90%的preferred样本的情况下，达到与DPO同样的性能。
3. 如果预训练模型足够好，相比于PPO和DPO必须在经过一次SFT，KTO可以直接跳过SFT阶段也能得到很好的性能。
4. 当数据噪声多、正负分布不均的情况下，KTO是一个很好的选择。但如果偏好对数据足够且噪声小，DPO或许表现更佳。

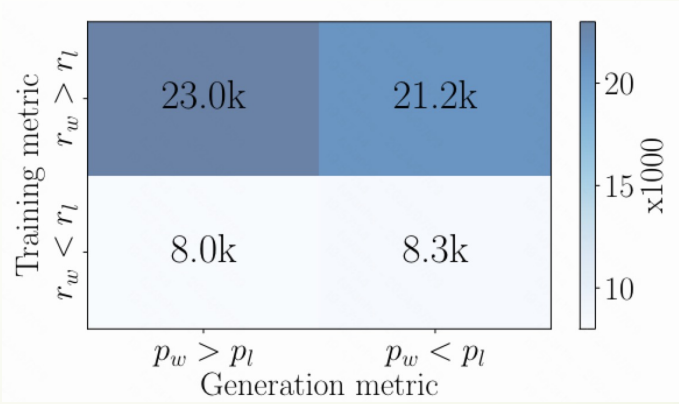
SimPO: Simple Preference Optimization with a Reference-Free Reward

- 使用 DPO 时，得到隐式奖励的方式是使用当前策略模型和参考模型之间的似然比的对数。但是，这种构建奖励的方式并未与引导生成的指标直接对齐，该指标大约是策略模型所生成响应的平均对数似然。
- 训练和推理之间的这种差异可能导致性能不佳。

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

$$p_\theta(y \mid x) = \frac{1}{|y|}\log\pi_\theta(y \mid x) = \frac{1}{|y|}\sum_{i=1}^{|y|}\log\pi_\theta(y_i \mid x, y_{<i}).$$

- 在DPO中，对于任意(x, y_w, y_l)，$r(x, y_w) > r(x, y_l)$ 并不意味着 $p_\theta(y_w \mid x) > p_\theta(y_l \mid x)$

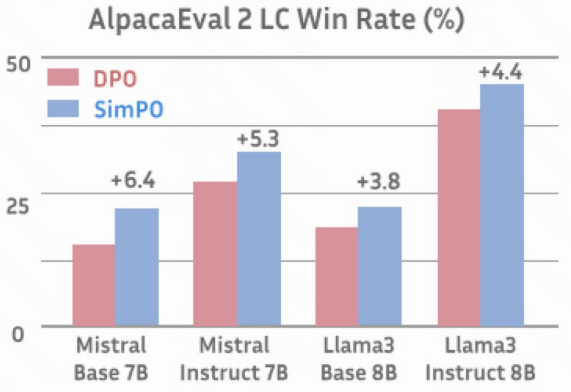


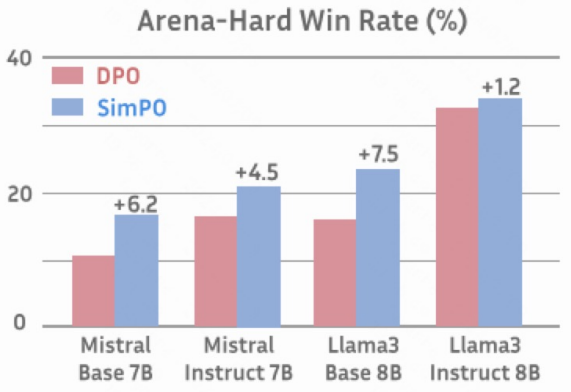
SimPO: Simple Preference Optimization with a Reference-Free Reward

- SimPO的核心是将偏好优化目标中的奖励函数与生成指标对齐。
- 包含两个主要组件：
  - （1）在长度上归一化的奖励；（引入是为了对齐生成的指标，但是后续实验发现移除这个因子会导致模型倾向于生成较长但质量较低的结果）
  - （2）目标奖励差额，用以确保chosen和reject之间的输出概率差保持一定差距；
  - 同时移除了参考模型，节约内存和计算效率

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

$$\mathcal{L}_{\text{SimPO}}(\pi_\theta) = -\mathbb{E}\left[\log \sigma\left(\frac{\beta}{|y_w|} \log \pi_\theta(y_w \mid x) - \frac{\beta}{|y_l|} \log \pi_\theta(y_l \mid x) - \gamma\right)\right]$$





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

SimPO: Simple Preference Optimization with a Reference-Free Reward

- 实验

Method	Mistral-Base (7B) Setting					Mistral-Instruct (7B) Setting				
	AlpacaEval 2		Arena-Hard	MT-Bench		AlpacaEval 2		Arena-Hard	MT-Bench	
	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4
DPO	15.1	12.5	10.4	5.9	7.3	26.8	24.9	16.3	6.3	7.6
SimPO	21.5	20.8	16.6	6.0	7.3	32.1	34.8	21.0	6.6	7.6
w/o LN	11.9	13.2	9.4	5.5	7.3	19.1	19.7	16.3	6.4	7.6
$\gamma = 0$	16.8	14.3	11.7	5.6	6.9	30.9	34.2	20.5	6.6	7.7

Arena-Hard数据集


```
urns":[{"content":"Write an SQL query to select the top 10 rows in a database and joins to 3 different table based
urns":[{"content":"I have a database table with columns account_id, day, balance. It holds the end-of-day balances
":[{"content":"How to sanitize inputs in argparse for Python to prevent special characters that can be used for SQ
":[{"content":"can you translate SQL \"SELECT * FROM SUBJECTS JOIN ON AUTHORS BY NAME\" to Datalog?\n"}]}
","turns":[{"content":"how can I use tailscale to securely expose a jellyfin server to the public internet?"}]}
","turns":[{"content":"Find root cause for this error:\nsshd[54785]: error: kex_exchange_identification: Connectio
urns":[{"content":"Create an \"impossible triangle\" with an SVG. Make it 3d"}]}
```


Method	Mistral-Base (7B)					Mistral-Instruct (7B)				
	AlpacaEval 2		Arena-Hard	MT-Bench		AlpacaEval 2		Arena-Hard	MT-Bench	
	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4
SFT	8.4	6.2	1.3	4.8	6.3	17.1	14.7	12.6	6.2	7.5
DPO [62]	15.1	12.5	10.4	5.9	7.3	26.8	24.9	16.3	6.3	7.6
IPO [6]	11.8	9.4	7.5	5.5	7.2	20.3	20.3	16.2	6.4	**7.8**
KTO [25]	13.1	9.1	5.6	5.4	7.0	24.5	23.6	17.9	6.4	7.7
ORPO [38]	14.7	12.2	7.0	5.8	7.3	24.5	24.9	20.8	6.4	7.7
R-DPO [60]	17.4	12.8	8.0	5.9	**7.4**	27.3	24.5	16.1	6.2	7.5
SimPO	**21.5**	**20.8**	**16.6**	**6.0**	7.3	**32.1**	**34.8**	**21.0**	**6.6**	7.6

Method	Llama3-Base (8B)					Llama3-Instruct (8B)				
	AlpacaEval 2		Arena-Hard	MT-Bench		AlpacaEval 2		Arena-Hard	MT-Bench	
	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4
SFT	6.2	4.6	3.3	5.2	6.6	26.0	25.3	22.3	6.9	8.1
DPO [62]	18.2	15.5	15.9	6.5	7.7	40.3	37.9	32.6	**7.0**	8.0
IPO [6]	14.4	14.2	17.8	6.5	7.4	35.6	35.6	30.5	**7.0**	**8.3**
KTO [25]	14.2	12.4	12.5	6.3	**7.8**	33.1	31.8	26.4	6.9	8.2
ORPO [38]	12.2	10.6	10.8	6.1	7.6	28.5	27.4	25.8	6.8	8.0
R-DPO [60]	17.6	14.4	17.2	**6.6**	7.5	41.1	37.8	33.1	**7.0**	8.0
SimPO	**22.0**	**20.3**	**23.4**	**6.6**	7.7	**44.7**	**40.5**	**33.8**	**7.0**	8.0

著名方法对比

Method	Objective
DPO [62]	$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$
IPO [6]	$\left(\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}-\frac{1}{2\tau}\right)^2$
KTO [25]	$-\lambda_w\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-z_{\text{ref}}\right)+\lambda_l\sigma\left(z_{\text{ref}}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right),$ where $z_{\text{ref}}=\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\beta\text{KL}\left(\pi_\theta(y\|x)\|\pi_{\text{ref}}(y\|x)\right)\right]$
ORPO [38]	$-\log p_\theta(y_w\|x)-\lambda\log\sigma\left(\log\frac{p_\theta(y_w\|x)}{1-p_\theta(y_w\|x)}-\log\frac{p_\theta(y_l\|x)}{1-p_\theta(y_l\|x)}\right),$ where $p_\theta(y\|x)=\exp\left(\frac{1}{\|y\|}\log\pi_\theta(y\|x)\right)$
R-DPO [60]	$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}-(\alpha\|y_w\|-\alpha\|y_l\|)\right)$
SimPO	$-\log\sigma\left(\frac{\beta}{\|y_w\|}\log\pi_\theta(y_w\|x)-\frac{\beta}{\|y_l\|}\log\pi_\theta(y_l\|x)-\gamma\right)$

谢谢！