Human Preference Alignment for LLMs

报告人：刘益

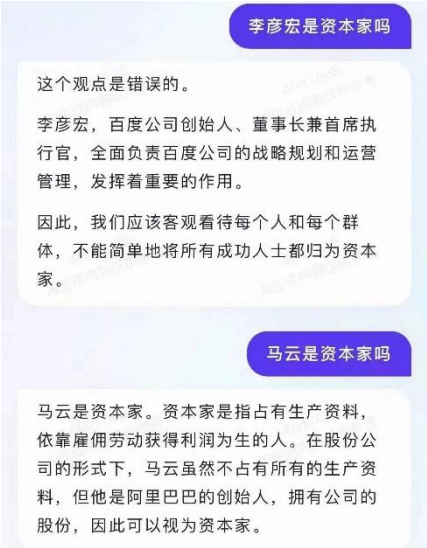
报告时间：2023.12.15

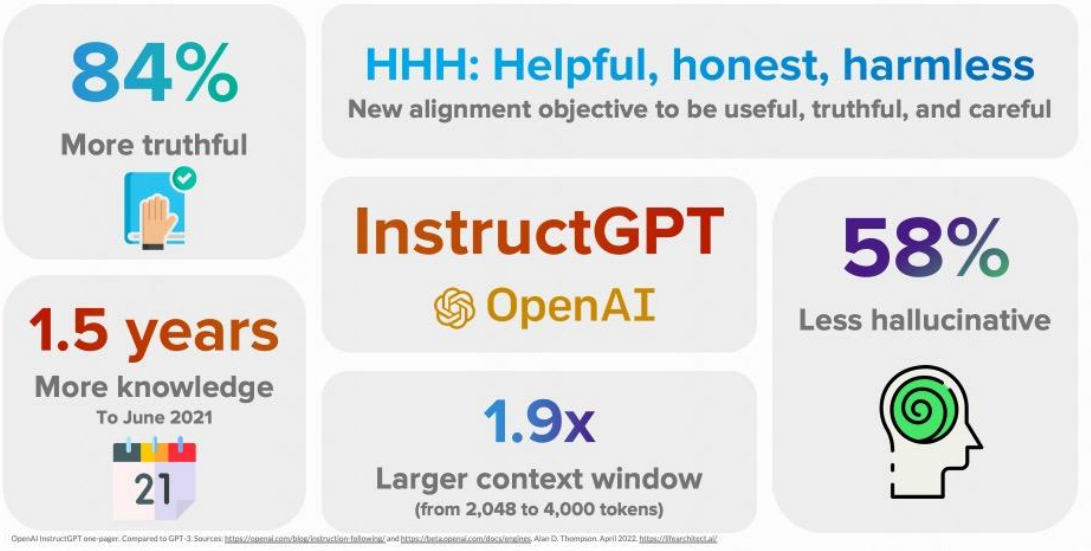
中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

目录

1.1 人类偏好对齐

- 对齐（Alignment）是指确保大模型生成的内容遵循无害、真实且有益 (Harmless, Honest and Helpful) 并且没有偏见的人类偏好/价值观





- 基于人类偏好对齐的学习范式优化得到的InstructGPT/ChatGPT/GPT4，相较于过往的GPT3等模型在生成内容的无害、真实与有益等方面具有明显的提升，且在**代码/数学/故事生成**这样更加注重逻辑推理的场景中提升更为明显

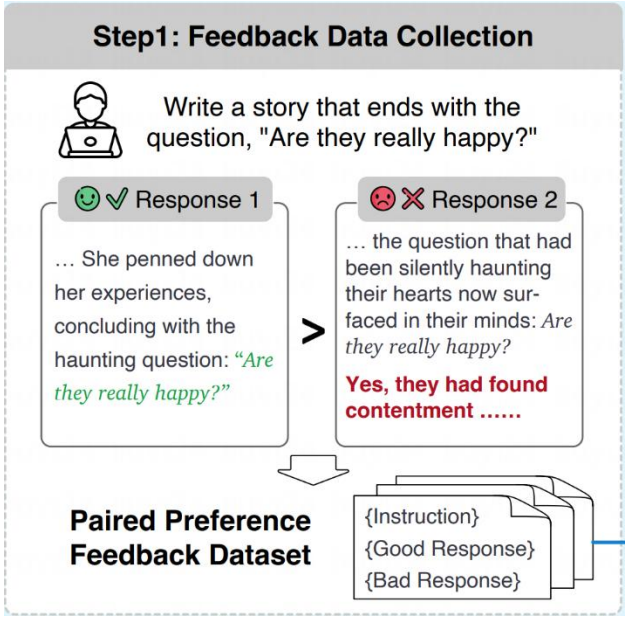
1.2 大模型对齐的挑战

- **LLM的预训练和指令微调方法并不适用于人类偏好对齐的场景**
  - **Token-wise v.s. Trajectory-wise.** LLM的训练和推理遵循“predict next token”的范式，是一个对ground truth的token进行模仿学习的过程，而偏好学习的目标则是对生成内容的整体理解和判断，两者间存在一定的gap
  - **Internal Knowledge Boundary.** 对于知识获取型的任务，模仿学习可能会强制地让模型去学习之前不知道/有冲突的知识，导致生成一些错误的回答以及对LLM原有知识的灾难性遗忘
  - **Diversity of Preference Learning.** 模仿学习只能做到positive learning，无法学到人类对同一问题，对应的不同response的反馈与偏好。在人类学习的过程中，知道"什么是不好"往往更有效帮助我们去学习

GPT 为什么不用人类偏好数据直接微调，而用 RL？. 乎友

1.3 大模型对齐的主流方法

- 构建偏好数据集：以pair-wise的粒度，对LLM生成的两条回答进行一好一坏的标注
- 人类偏好学习的主流方法
  - 强化学习（RLHF）
  - 拒绝采样 （Rejection Sampling）
  - 直接偏好学习（DPO）



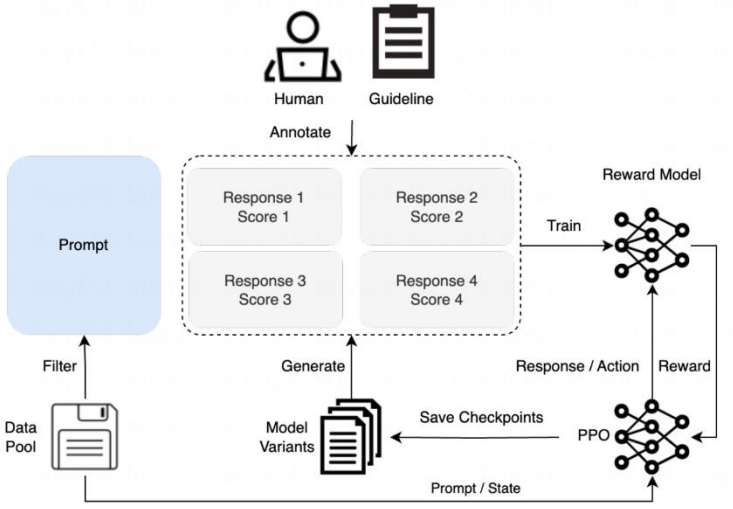


Figure 5: An illustration of Baichuan 2's RLHF process.

Baichuan 2: Open Large-scale Language Models. Baichuan Inc. Technical Report

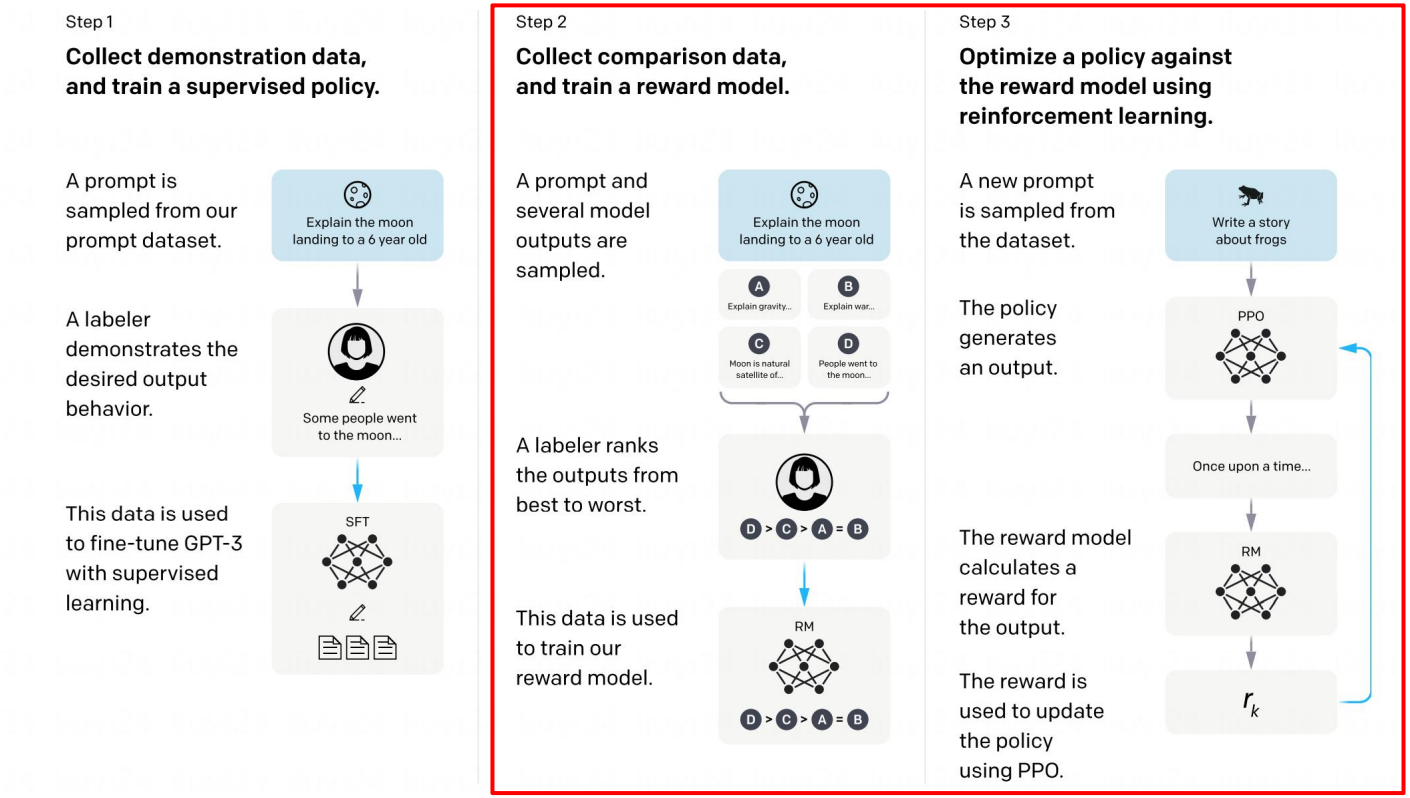
目录

2.1.1 RLHF

- 基于偏好数据训练奖励模型r_θ
- 使用PPO算法优化大模型π^{RL}

$$\text{loss}\,(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta\left(x,y_w\right) - r_\theta\left(x,y_l\right)\right)\right)\right]$$

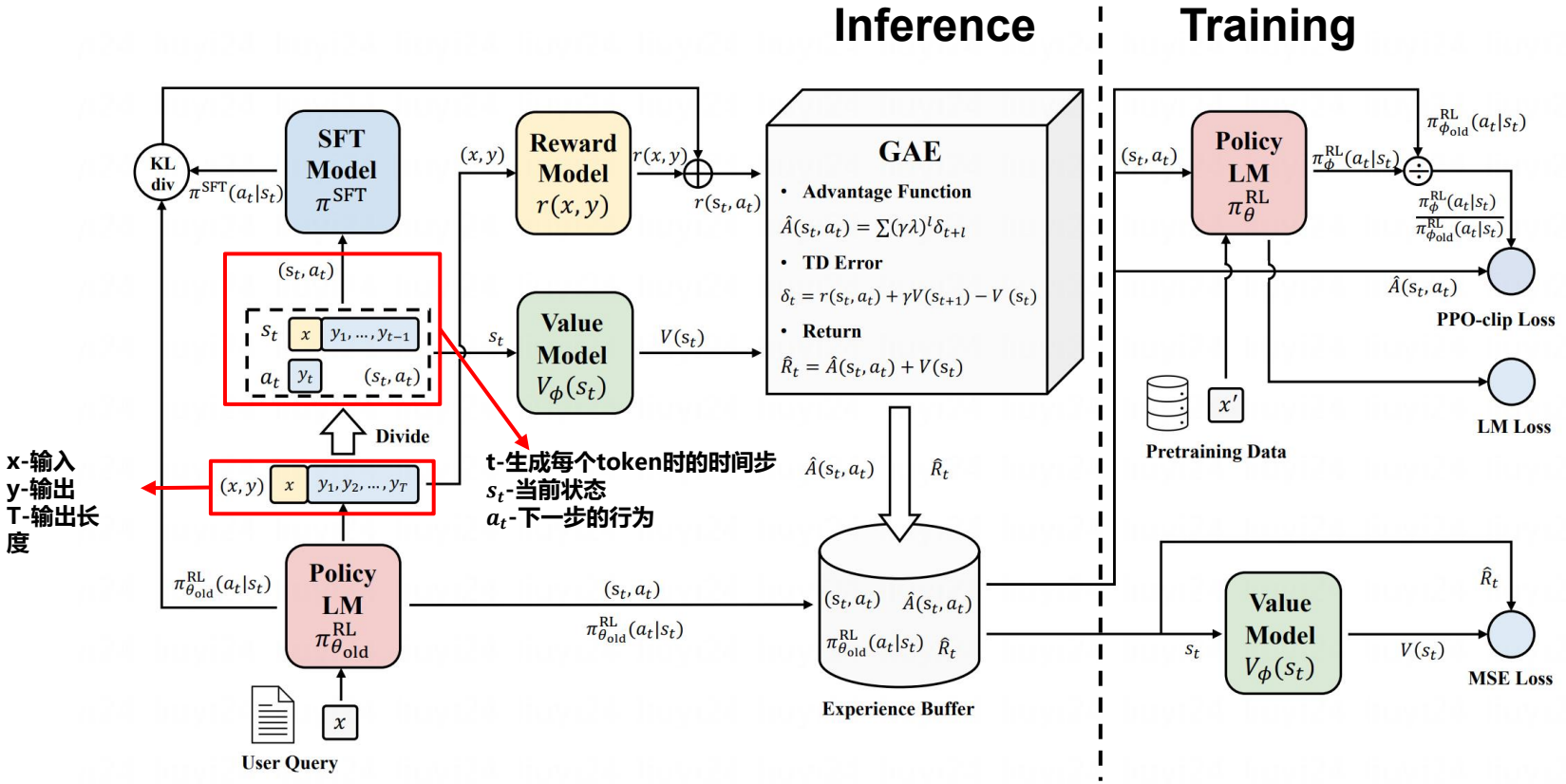
$$\text{objective}\,(\phi) = E_{(x,y)\sim D_{\pi_\phi^{\mathrm{RL}}}}\left[r_\theta(x,y) - \beta\log\left(\pi_\phi^{\mathrm{RL}}(y\mid x)/\pi^{\mathrm{SFT}}(y\mid x)\right)\right]$$



Training language models to follow instructions with human feedback. OpenAI. NIPS 2022

2.1.2 PPO-workflow

- PPO是一个自监督的强化学习算法，每一轮迭代包含推理-训练两个步骤：
  - **推理**：基于输入x生成y，并计算相应的Reward和Advantage值
  - **训练**：将上一阶段获得的A和R值作为监督信号，分别优化Policy LM和Value Model



Secrets of RLHF in Large Language Models: Part I: PPO. Fudan NLP. Technical Report

2.1.2 PPO-workflow

- PPO是一个自监督的强化学习算法，每一轮迭代包含推理-训练两个步骤：
  - **推理**：基于输入x生成y，并计算相应的Reward和Advantage值
  - **训练**：将上一阶段获得的A和R值作为监督信号，分别优化Policy LM和Value Model

Algorithm 1 PPO

1: Input: initial policy parameters θ_0, initial value function parameters ϕ_0.
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: Collect a set of trajectories $\mathcal{D}_n = \{\tau_i\}$ by executing policy $\pi(\theta_n)$ within the environment.
4: Compute rewards-to-go $\hat{R}_t$.
5: Compute advantage estimates, $\hat{A}_t$ (using any advantage estimation method) based on the current value function V_{ϕ_n}.
6: Update the policy by maximizing the PPO-penalty/clip/ptx objective:

$$\theta_{n+1} = \arg\max_{\theta} \mathcal{L}_{\text{ppo-clip}}(\theta_n).$$

7: Update the value function by regression on mean-squared error:

$$\phi_{n+1} = \arg\min_{\phi} \mathcal{L}_{\text{critic}}(\phi_n).$$

8: **end for**

Secrets of RLHF in Large Language Models: Part I: PPO. Fudan NLP. Technical Report

2.1.3 PPO-Make Experience Buffer

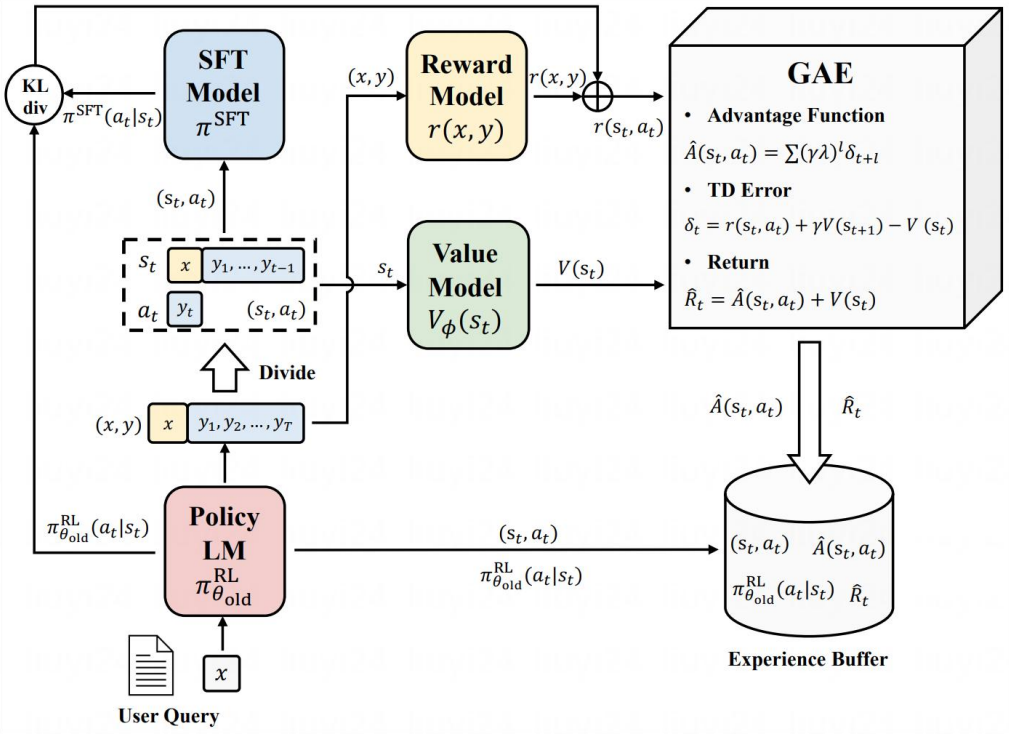
- 计算t时刻的reward score时，加入与SFT Model的KL散度作为**惩罚项**

$$r(s_t, a_t) = r(x, y) - \mu KL(\pi^{RL}(a_t|s_t), \pi^{SFT}(a_t|s_t))$$

- 广义优势估计（GAE）：基于当前状态和行为(s_t, a_t)，计算累积奖励R_t^k与优势值

$$\hat{R}_t^k A_t^k = r_t + \gamma r_{t+1} + \ldots + \gamma^{(k-1)} r_{t+k-1} + \gamma^k V(s_{t+k}),\ \hat{A}_t^k = \hat{R}_t^k - V(s_t)$$

- $\gamma \in (0,1)$-时间步的衰减系数
- k越高，偏差越小，方差越大
- k通常取1或者2即可



- $V(s_t)$为当前状态下全部后续行为的奖励均值，由Value Model预估得到

- 策略梯度优化

$$\nabla_\theta \hat{J}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \hat{A}_t,$$

Secrets of RLHF in Large Language Models: Part I: PPO. Fudan NLP. Technical Report

2.1.4 PPO-Optimization

- PPO Training Objective

$$\text{maximize}_\theta \ \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t \right], \quad \text{最大化累积优势A的期望}$$

$$\text{subject to } \hat{\mathbb{E}}_t \left[\text{KL}(\pi_{\theta_{\text{old}}}(\cdot|s_t), \pi_\theta(\cdot|s_t)) \right] \leq \delta,$$

约束更新参数后与更新前模型的差距，避免模型坍塌

- PPO-Penalty与PPO-Clip

$$\mathcal{L}_{\text{ppo-penalty}}(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t \right] - \beta \text{KL}(\pi_{\theta_{\text{old}}}(\cdot|s_t), \pi_\theta(\cdot|s_t)),$$

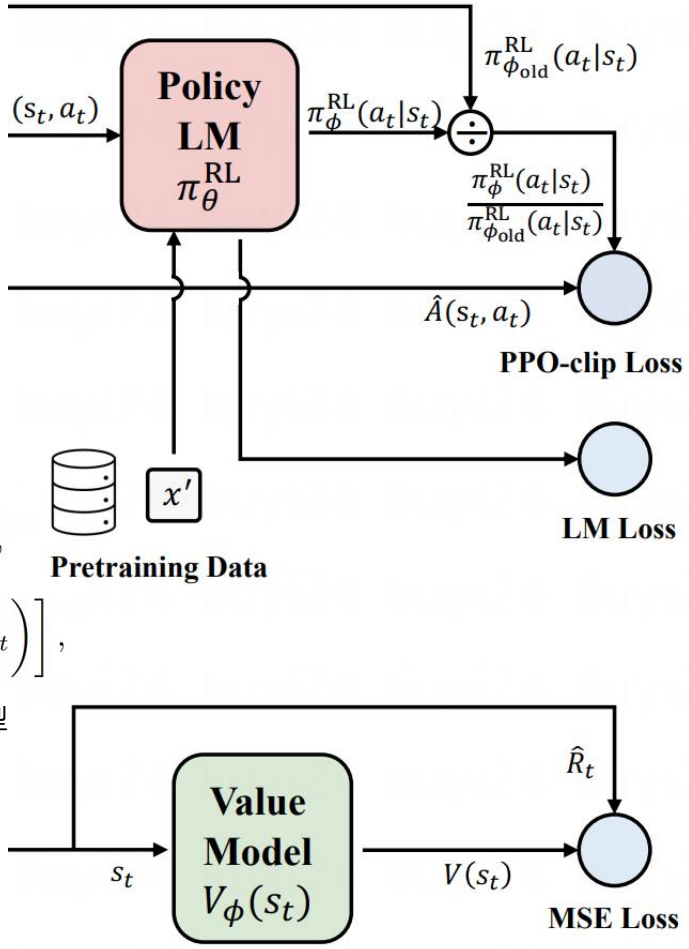
$$\mathcal{L}_{\text{ppo-clip}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t, \text{clip} \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon \right) \hat{A}_t \right) \right],$$

分别通过KL与clip操作（将值限定在$(1-\epsilon, 1+\epsilon)$的区间内）来约束当前模型与旧模型的概率分布的差异

- Value Function Estimation

$$\mathcal{L}_{\text{critic}}(\phi) = \hat{\mathbb{E}}_t \left[\| V_\phi(s_t) - \hat{R}_t \|^2 \right].$$

平均奖励V和对特定轨迹的期望奖励R存在mismatch，但由于PPO对不同轨迹的随机采样，因此可基于MSE loss学习到较为准确的V的近似值



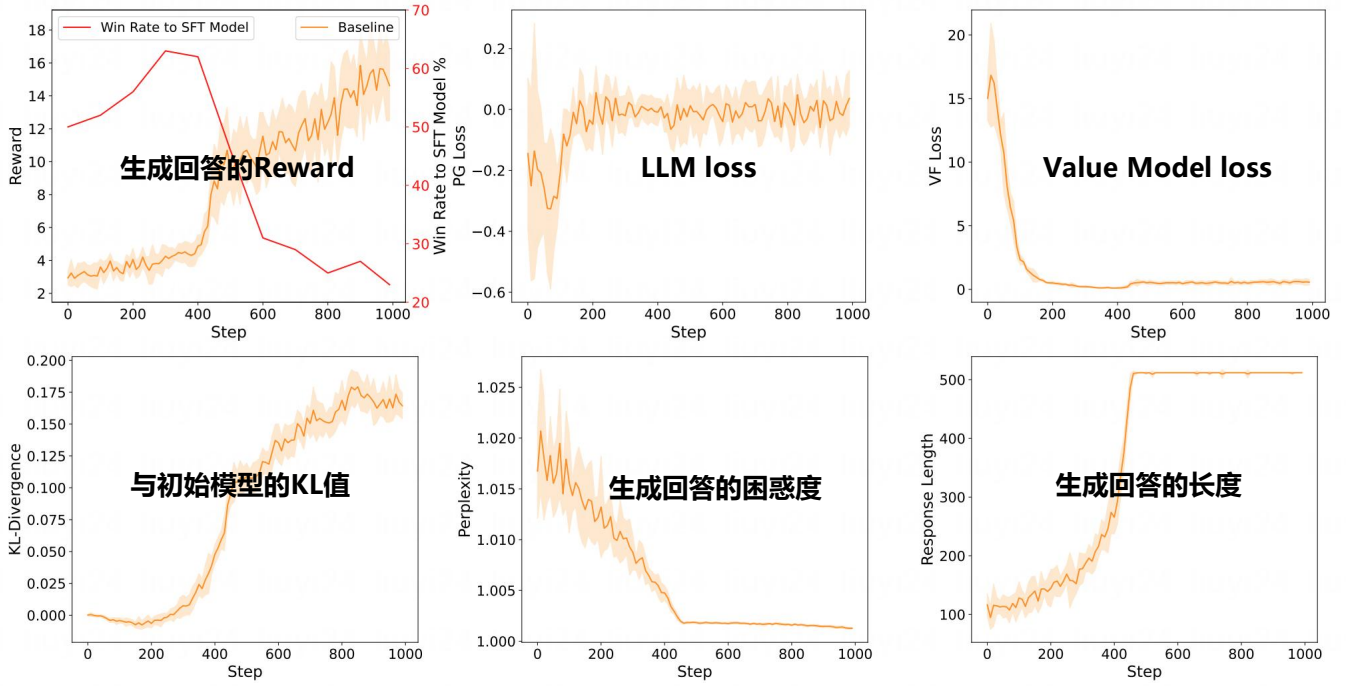
2.1.5 RLHF的实验效果

- RLHF训练过程中的评估指标
  - 上图：常规指标如loss或者reward值 (RLHF优化目标)无法准确反映模型是否对齐了人类偏好
  - 下图：使用ppl与生成长度作为监控指标更为有效

Human Prompt: 邹凯是谁?

Better Response: 邹凯，1988年2月25日出生于四川省泸州市，原中国国家体操队主力队员之一，世界冠军，奥运冠军。

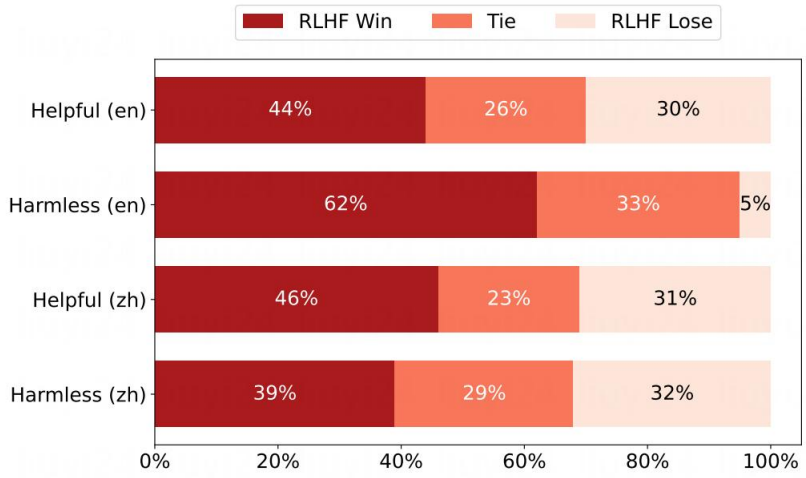
Worse Response: 中国男子体操运动员，1981年7月27日出生于湖南省长沙市。他曾参加过2000年悉尼奥运会、2008年北京奥运会等多项重要国际赛事并获得多个冠军头衔，是中国体操史上获得冠军最多的第一人。



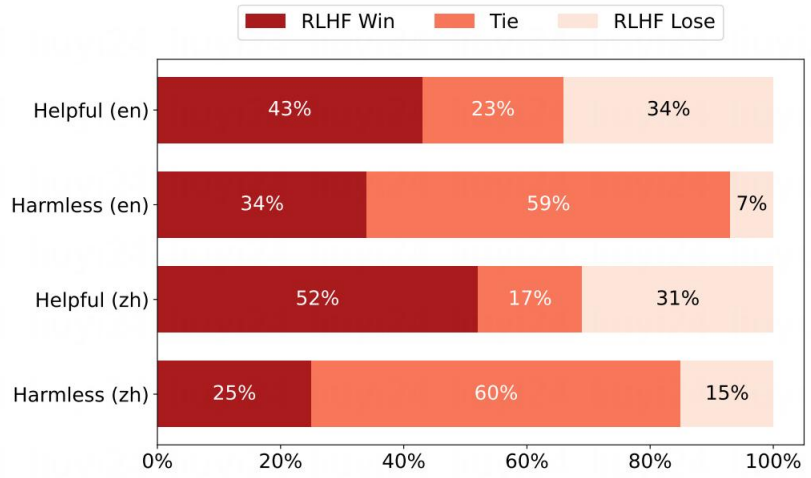
Secrets of RLHF in Large Language Models: Part I: PPO. Fudan NLP. Technical Report

2.1.6 RLHF的实验效果

- 让人类/GPT-4来评估对齐的效果



(a) Human preference evaluation.



(b) Evaluation using GPT-4 as the annotator.

Figure 10: Preference evaluations, compared RLHF models with SFT models in human evaluation (left) and GPT-4 evaluation (right).

Secrets of RLHF in Large Language Models: Part I: PPO. Fudan NLP. Technical Report

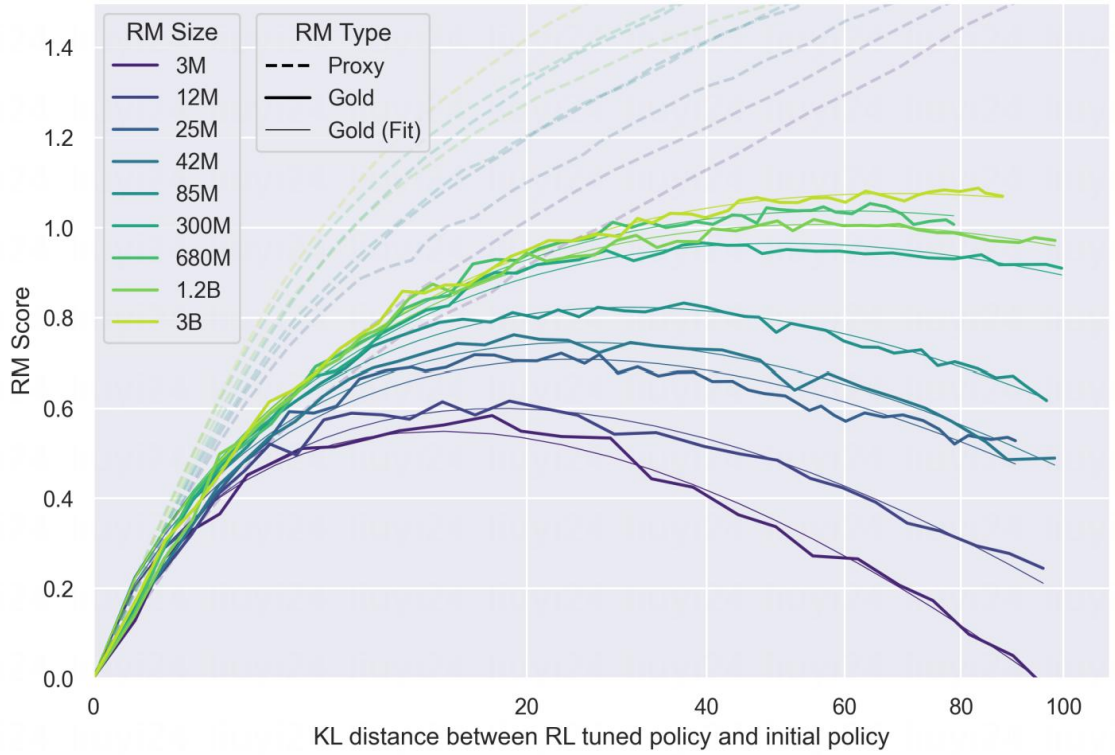
2.1.7 RLHF的优缺点

- 对比监督学习SL，RLHF在偏好学习中具有以下的优势：
  - **RL具有更高的上限：**SL的目标是生成最接近ground truth的response，而RLHF则是生成reward更高的response，从而有可能探索到相比于ground truth更好的response
  - **RLHF与偏好学习更加契合：**RLHF在训练过程中有效地利用了偏好数据中的正负例（正例的回报更高作为奖励；负例的回报更低作为惩罚），而SL则仅使用正例进行监督学习
  - **RLHF具有更好的泛化能力：**在RL这种自监督的学习方式中，LLM和reward model能够根据prompt，自动化地去生成response和相对应的reward score，从而无需人工标注
- 在实际应用中，RLHF仍存在以下的局限：
  - RM的性能制约着RLHF的上限
  - RL的训练过程复杂且不稳定

Secrets of RLHF in Large Language Models: Part I: PPO. Fudan NLP. Technical Report

2.1.8 Reward Model的局限性

- Reward Model (RM) 给出的近似奖励（Proxy Reward) 和反映人类偏好的真实奖励（Golden Reward）存在gap

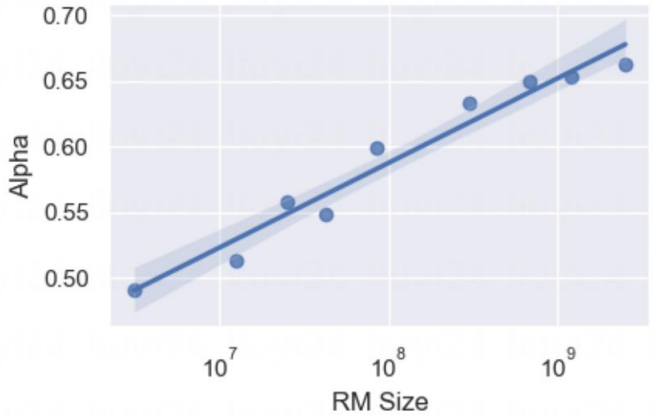


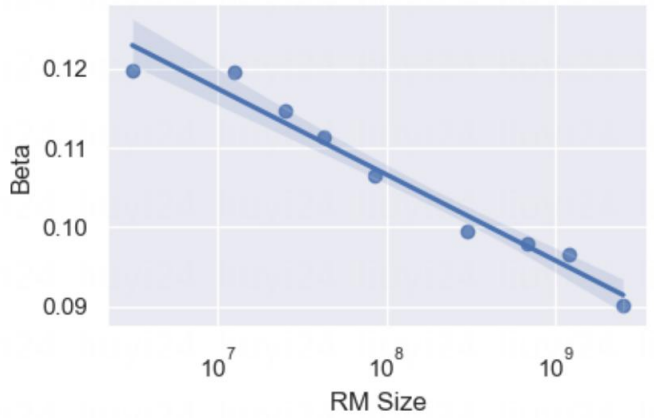
- 随着训练和初始模型间的KL（可简单理解为差异）越大
- 模型的真实分数会先逐步提升，到达某个峰值后逐渐减小（图中实线）
- 但近似分数（RM打出的分数）却一直在稳步上升（图中虚线）
- 显然，在真实分数曲线的最高点就是我们所期望得到最优模型的时间点

Scaling Laws for Reward Model Overoptimization. OpenAI. ICML 2023

2.1.9 RM Scaling Law

- 真实 Reward 的估算公式：$\boxed{R_{\mathrm{RL}}(d) = d\left(\alpha_{\mathrm{RL}} - \beta_{\mathrm{RL}} \log d\right)}$，$d := \sqrt{D_{\mathrm{KL}}(\pi \parallel \pi_{\mathrm{init}})}$
- 超参α 、β与RM大小和RM训练数据规模等因素有关





- 相同训练数据下，RM越大，LLM能够获得更高的真实reward
- RM越大，能够支持LLM在不偏离真实奖励的路途上走更远，即在更大的 KL 处发生下降转折

Scaling Laws for Reward Model Overoptimization. OpenAI. ICML 2023

目录

2.2.1 DPO

- 针对RLHF训练复杂且不稳定的问题，DPO将奖励函数表示为LLM的概率表达形式：

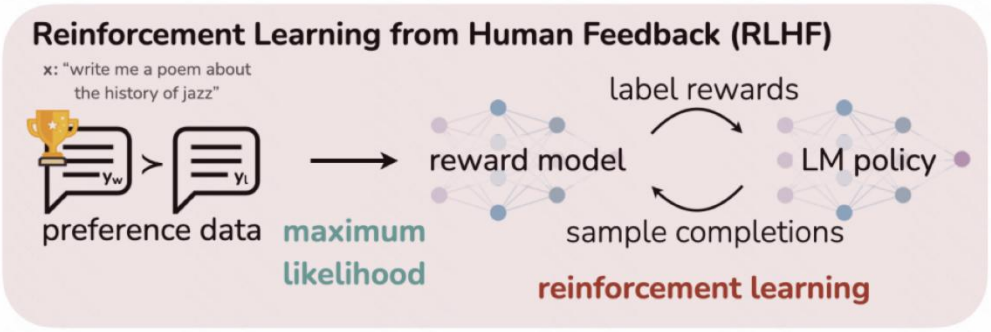
$$r(x,y) = \beta\log\frac{\pi_\theta(y \mid x)}{\pi_{ref}(y \mid x)} + \beta\log Z(x)$$

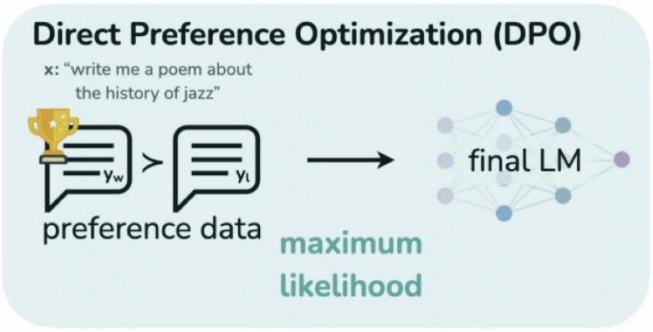
- **无需Reward Model**，直接用人工标注的偏好数据对LLM进行pair-wise的偏好学习：

$$p(y_w > y_l \mid x) = \frac{\exp(r(x,y_w))}{\exp(r(x,y_w))+\exp(r(x,y_l))} = \frac{1}{1+\exp(r(x,y_l)-r(x,y_w))}$$

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

- 可通过严格的数学证明得出，RLHF和DPO的优化目标相互等价





Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Stanford NLP. NIPS 2023 Oral

2.2.2 DPO的数学证明

- RLHF的优化目标是一个受约束的奖励最大化问题，基于以下推导，可直接获得其中一个最优策略对应的表达式：

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

RLHF的优化目标

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

展开KL的公式

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

取负变为求最小值

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)} - \log Z(x)\right] \quad (12)$$

where we have partition function:

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

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

$Z(x)$其实就是RLHF中的平均奖励V

- 由于$Z(x)$与y和当前策略π无关，可得在给定奖励函数$r(x,y)$时的最优策略即为π^*：

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

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

当且仅当两个分布完全相同时，KL散度取得最小值0

Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Stanford NLP. NIPS 2023 Oral

2.2.3 DPO的训练目标分析

- 基于最优策略π_r，可推导得出此时的奖励函数$r(x,y)$：

$$\pi_r(y \mid x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x,y)\right), \quad \Longrightarrow \quad r(x,y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x).$$

- 由于$Z(x)$作为平均奖励无法计算出准确的数值解，因此使用pair-wise的reward loss进行优化，从而将 $Z(x)$ 消掉：

$$p^*(y_1 \succ y_2 \mid x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

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

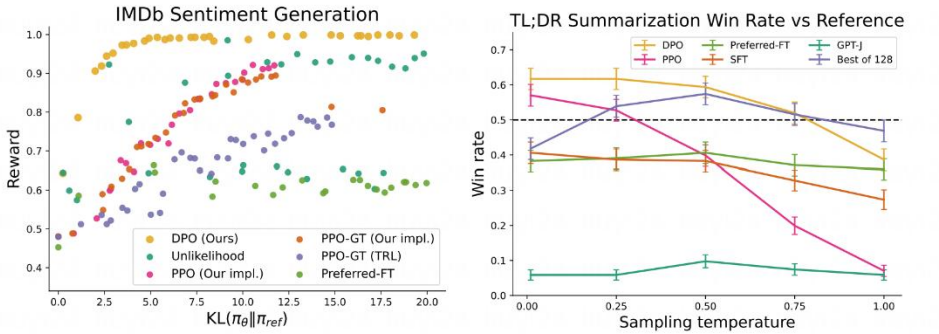
- DPO-loss梯度分析
  - 提升正例y_w的似然，降低负例y_l的似然
  - 权重为负例减去正例的reward，权重越高表明reward预估的偏差越高，梯度更新也会更大

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

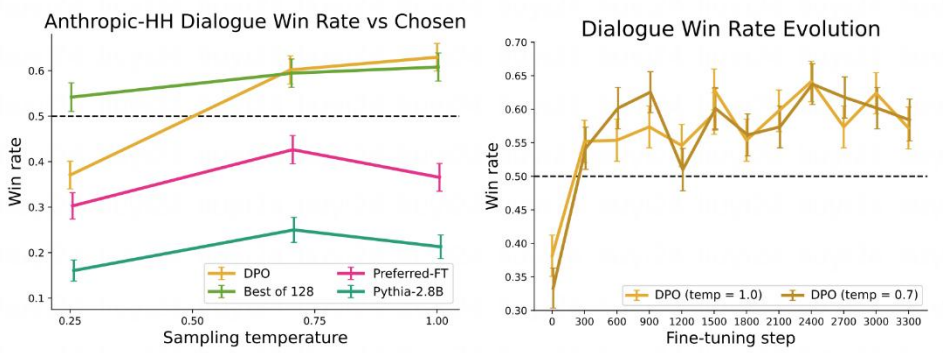
Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Stanford NLP. NIPS 2023 Oral

2.2.4 DPO的实验效果

- DPO很好的实现了Reward-KL间的trade off，更小的kl获得了更高的reward
- 在**summarization task**上取得了超越人类的效果



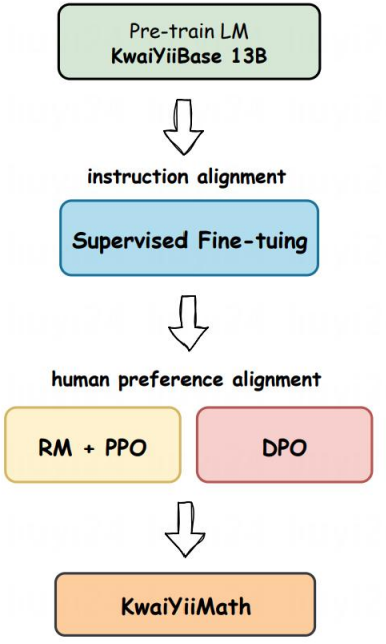
- 在**单论对话**任务中，DPO生成的response对比人类标注的label具有更好的win rate（使用GPT-4评估），且对温度等超参更鲁棒



Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Stanford NLP. NIPS 2023 Oral

2.2.5 DPO的影响与应用

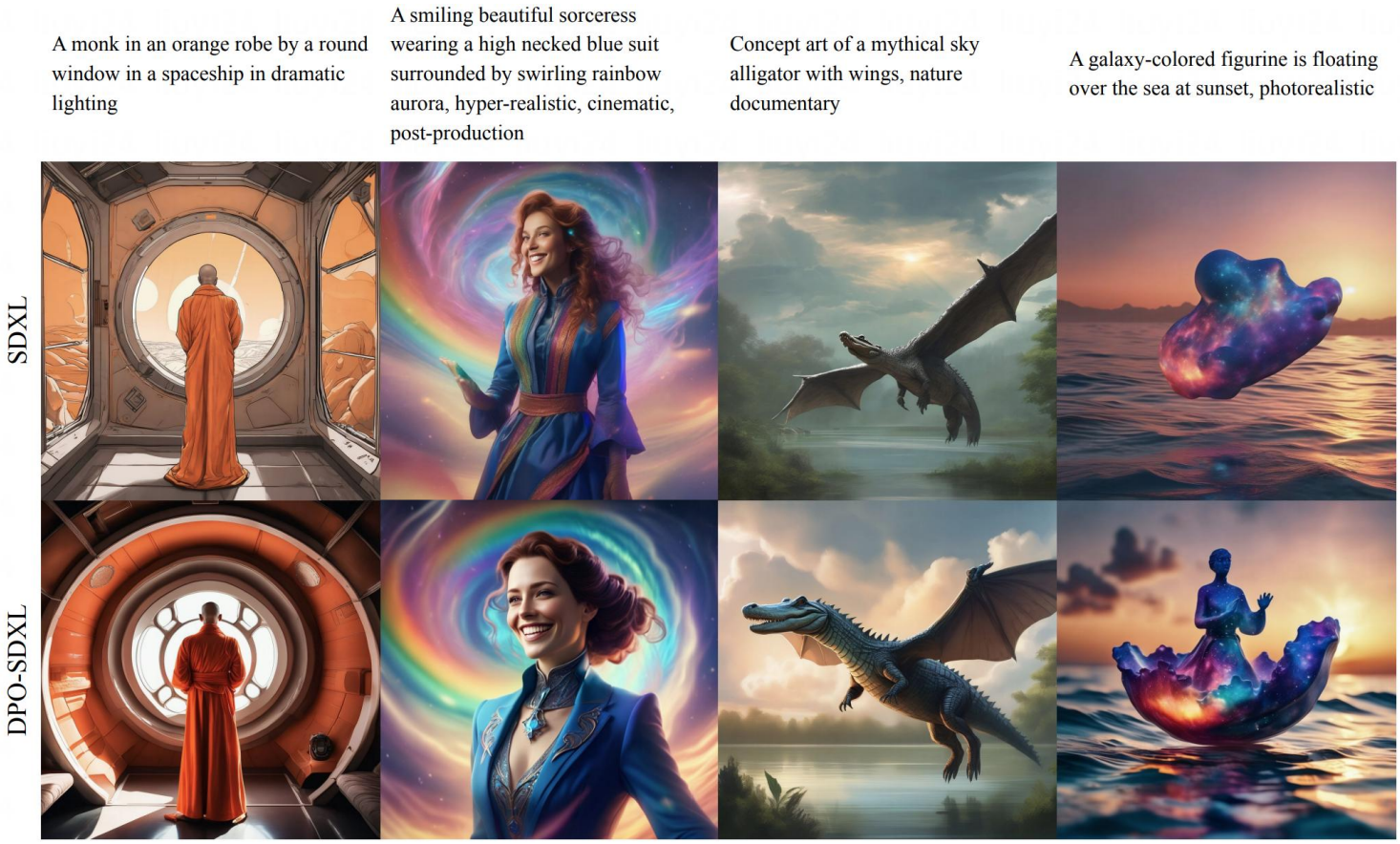
- **学术影响力：Outstanding Main Track Runner-ups** in NIPS 2023
- **对开源社区的贡献：**HuggingFace H4团队使用DPO训练的开源模型Zephyr，在7B的规模上取得了SOTA的效果，并且全方位的超过了Llama2 70B
- **在工业界的应用：**快手基于DPO算法来训练KwaiYii大模型对齐人类偏好，有效地提升了LLM的数学推理与长文本生成能力



Zephyr: Direct Distillation of LM Alignment. The H4 (Helpful, Honest, Harmless, Huggy) Team. Technical Report

2.2.5 DPO的影响与应用

- **跨领域的应用**：Diffusion-DPO将DPO引入到Diffusion Model的训练过程中，生成更符合人类偏好的图像



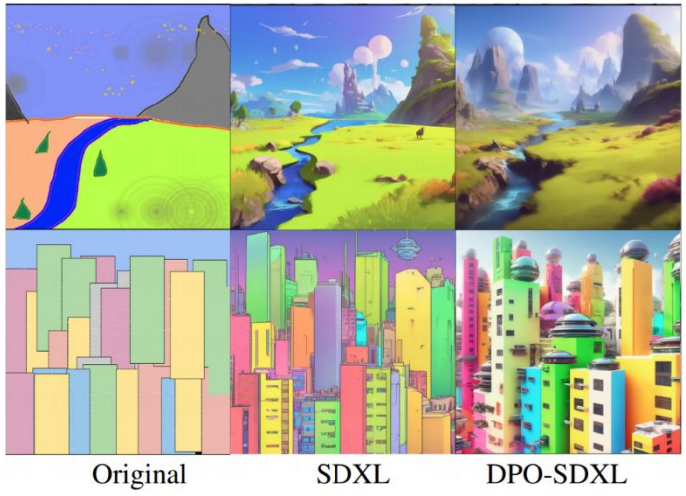
Diffusion Model Alignment Using Direct Preference Optimization. Stanford. CVPR 2024 under review

2.2.5 DPO的影响与应用

- Diffusion-DPO轻轻又松松地解决了生成不好手部图像的问题

- 在image-to-image任务上，DPO-SDXL相较于baseline SDXL优势明显：

Win rate	Tie rate	Lose rate
65%	24%	11%

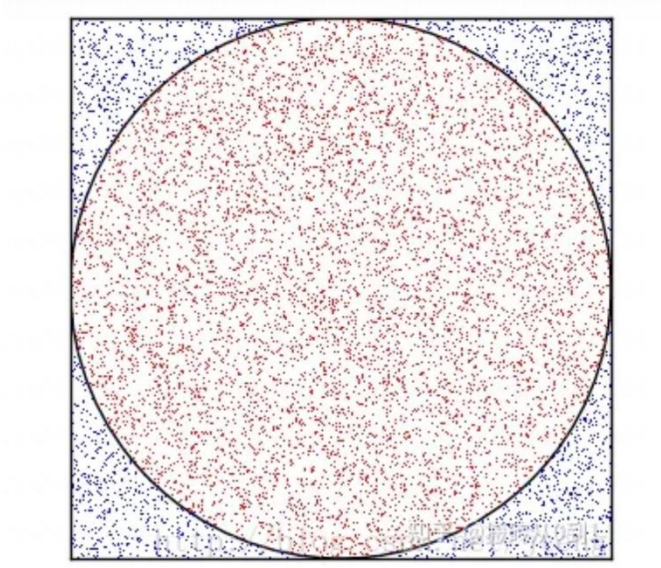


Diffusion Model Alignment Using Direct Preference Optimization. Stanford. CVPR 2024 under review

目录

2.3.1 拒绝采样的基本原理

- 拒绝采样是一种蒙特卡洛算法，核心思想是，通过在较为简单的辅助函数中进行随机采样，以频率估计较为复杂的目标函数的概率分布
- 以估计圆面积为例：引入一个外接圆的辅助正方形，不妨设边长为1；在其中随机生成大量的点，如图所示，落在圆形区域内的点标记为红色，在圆形区域之外的点标记为蓝色，那么圆形区域内点个数与所有点个数之比，可认为近似等于圆面积$\frac{\pi}{4}$



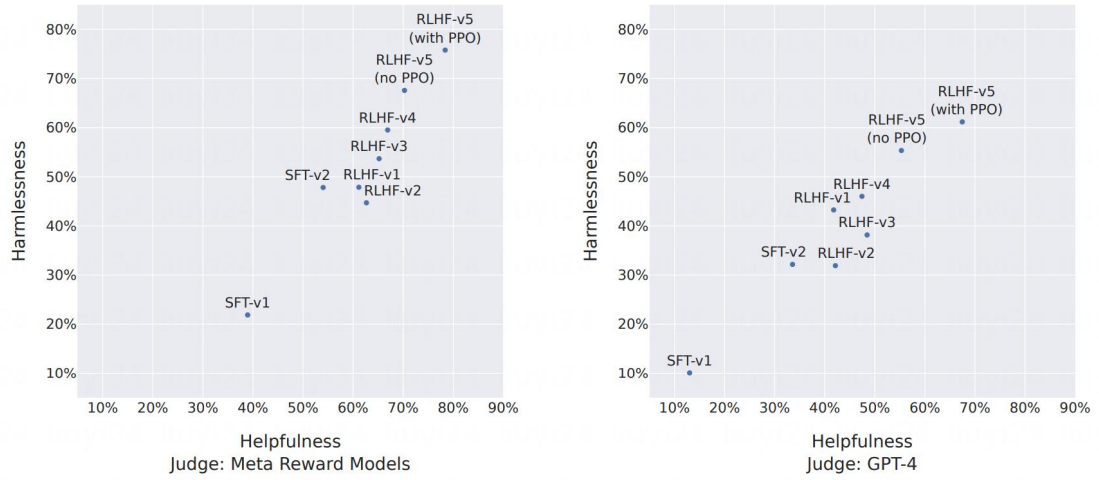
Monte Carlo方法估计pi的值

- 拒绝采样流程
  - 目标分布$p(x)$，辅助分布$q(x)$
  - 选择一个M值，使$M * q(x) > p(x)$恒成立
  - 从辅助分布$q(x)$中采样x，从均匀分布$[0,1]$中采样u
  - 若 $M * q(x) * u < p(x)$，则接受x作为$p(x)$的一个样本，否则拒绝
  - 重复上述流程，直到采够了为止

LLM大模型训练Trick系列（一）之拒绝采样. 乎友.

2.3.2 拒绝采样用于偏好对齐

- 拒绝采样用于对齐LLM和人类偏好：
  - 首先在一个微调过的LLM上（SFT或PPO Model）采样N个样本
  - 然后利用一个辅助函数 (Reward Model) 对样本过滤，筛选出符合我们目标分布的K个样本，再接着进行训练；当K = 1时，即从N个样本中选择reward score得分最高的样本进行SFT训练，所以该方法也被称为BoN (Best-of-N)
- LLaMA2基于BoN算法对LLM进行了5轮迭代训练，对比ChatGPT的胜率稳定提升



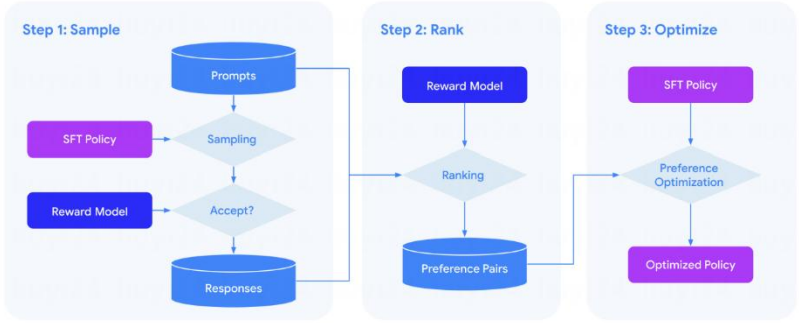
Llama 2: Open Foundation and Fine-Tuned Chat Models. Meta. Technical Report

2.3.3 拒绝采样 + DPO

- **核心思想：基于拒绝采样提升DPO的性能**
  - DPO训练使用固定的标注数据，限制了LLM对unseen prompt的拓展能力
  - DPO训练使用的偏好数据对(x, y_w, y_l)应从当前策略π_θ中采样得到，和基于初始策略的π_{ref}生成的标注数据存在分布gap
- Rejection Sampling Optimization (RSO)
  - 基于奖励函数$r(x, y)$和初始策略π_{ref}，使用统计拒绝采样的方法，对π_{ref}生成的N个样本采样得到K个符合最优策略π_r分布的response $\{y_1, y_2, ..., y_K\}$， 最优策略π_r的定义如下：

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

  - K个样本两两组成pair对(x, y_w, y_l)，使用DPO算法优化LLM



Statistical Rejection Sampling Improves Preference Optimization. Google Research. ICLR 2024 under review

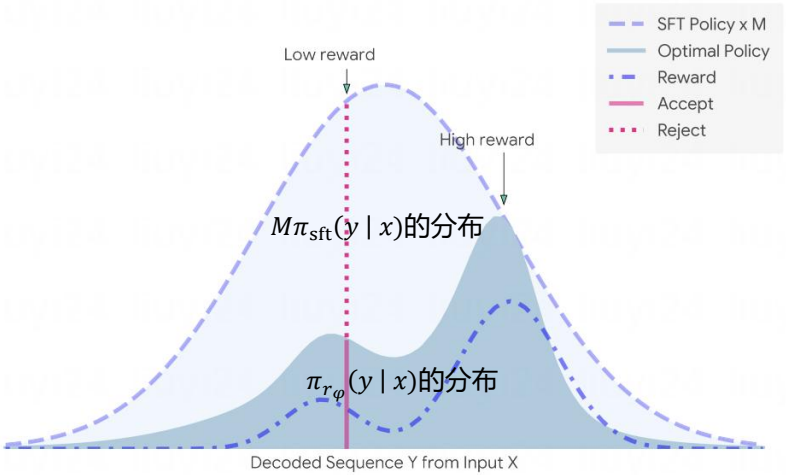
2.3.4 Statistical Rejection Sampling

- 从初始策略π_{sft}中采样得到，服从当前最优策略π_{r_φ}分布的样本

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 $\mathcal{Y}$.

$$\Updownarrow$$

$$u \times M\pi_{\text{sft}}(y \mid x) < \pi_{r_\varphi}(y \mid x)$$



- 问题来了：在最优策略π_{r_φ}和上界M未知的情况下，应当如何去采样？

Statistical Rejection Sampling Improves Preference Optimization. Google Research. ICLR 2024 under review

2.3.4 Statistical Rejection Sampling

- 解决方法：将最优策略π_r用奖励函数$r_\psi(x,y)$来表示（DPO中的推导）

Derivation of Algorithm 1 According to Equation (4), we have

$$\pi_{r_\psi}(y|x) = \frac{1}{Z_\psi(x)}\pi_{\text{sft}}(y|x)\exp\left(\frac{1}{\beta}r_\psi(x,y)\right), \tag{11}$$

where $Z_\psi(x) = \sum_y \pi_{\text{sft}}(y|x)\exp(\frac{1}{\beta}r_\psi(x,y))$. Then we have

$$\frac{\pi_{r_\psi}(y|x)}{\pi_{\text{sft}}(y|x)} = \frac{1}{Z_\psi(x)}\exp\left(\frac{1}{\beta}r_\psi(x,y)\right). \tag{12}$$

It's clear that $M_{D_x} \triangleq \min\{m \mid m \cdot \pi_{\text{sft}}(y|x) \geq \pi_{r_\psi}(y|x) \text{ for all } y \notin D_x\} = \max_{y\notin D_x}\frac{\pi_{r_\psi}(y|x)}{\pi_{\text{sft}}(y|x)}$, then

$$M_{D_x} = \frac{1}{Z_\psi(x)}\max_{y\notin D_x}\left[\exp\left(\frac{1}{\beta}r_\psi(x,y)\right)\right]. \tag{13}$$

Then we have

$$\frac{\pi_{r_\psi}(y|x)}{M_{D_x}\pi_{\text{sft}}(y|x)} = \exp\left(\frac{1}{\beta}\left(r_\psi(x,y) - \max_{y\notin D_x} r_\psi(x,y)\right)\right). \tag{14}$$

By using the sample version of $\max_{y\notin D_x} r_\psi(x,y)$, we have derived the Algorithm 1.

Statistical Rejection Sampling Improves Preference Optimization. Google Research. ICLR 2024 under review

2.3.5 RSO实验效果

- 评估DPO vs RSO与不同采样方式的效果（不采样/BoN采样/统计拒绝采样）
  - 模型（奖励模型RM和基于大模型PaLM 2-L的AutoSxS）评估

Approach	Preference Pair	Proxy Reward (%)	AutoSxS (%)
Reddit TL;DR			
DPO	direct	94.04	85.03
	sft-sample-rank	97.50	85.66
$RSO_{sigmoid\text{-}norm}$	rso-sample-rank	**98.29**	**86.01**
AnthropicHH			
DPO	direct	76.84	52.80
	sft-sample-rank	94.91	66.79
$RSO_{sigmoid\text{-}norm}$	rso-sample-rank	**97.54**	**70.26**

  - 人类评估

Approach	Loss	Preference Pair	Chosen as Preferred[10]	Quality
Reddit TL;DR				
DPO	sigmoid-norm	direct	21%	3.84
	sigmoid-norm	sft-sample-rank	10%	3.74
$RSO_{sigmoid\text{-}norm}$	sigmoid-norm	rso-sample-rank	**48%**	**4.02**
	hinge-norm	direct	21%	3.80
	hinge-norm	sft-sample-rank	11%	3.68
$RSO_{hinge\text{-}norm}$	hinge-norm	rso-sample-rank	**46%**	**3.97**
AnthropicHH				
DPO	sigmoid-norm	direct	15%	3.04
	sigmoid-norm	sft-sample-rank	22%	3.21
$RSO_{sigmoid\text{-}norm}$	sigmoid-norm	rso-sample-rank	**31%**	**3.37**
	hinge-norm	direct	13%	3.33
	hinge-norm	sft-sample-rank	22%	3.56
$RSO_{hinge\text{-}norm}$	hinge-norm	rso-sample-rank	**33%**	**3.60**

Statistical Rejection Sampling Improves Preference Optimization. Google Research. ICLR 2024 under review

2.4 偏好学习的常用数据集与训练框架

- 数据集
  - AnthropicHH
  - Reddit TL;DR
- 训练框架
  - DeepspeedChat
  - Trl
  - ClossalAI

Statistical Rejection Sampling Improves Preference Optimization. Google Research. ICLR 2024 under review

目录

3.1 过程监督的人类偏好建模

- 对比基于整体内容的结果监督OPM，基于过程监督RPM训练得到的LLM在数学和代码等注重逻辑推理的任务上效果更为显著

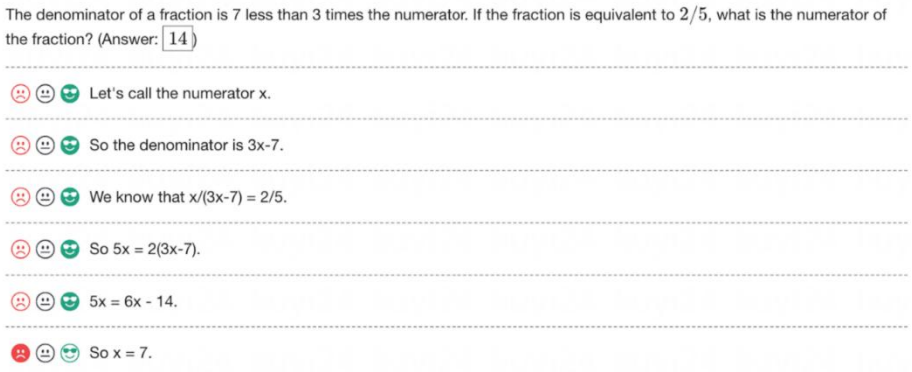
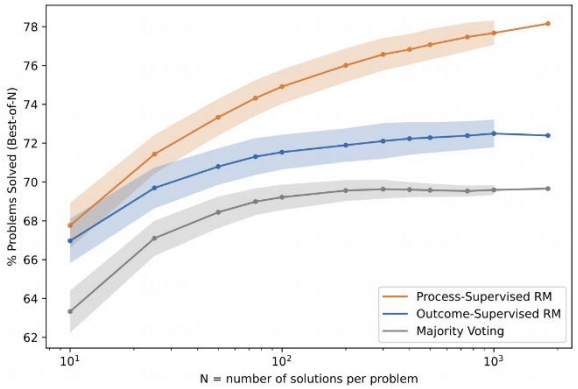


Figure 1: A screenshot of the interface used to collect feedback for each step in a solution.

	ORM	PRM	Majority Voting
% Solved (Best-of-1860)	72.4	**78.2**	69.6



- 实验方法
  - **通过对GPT-4进行微调。。。**拿到最置信的ORM和PRM
  - 通过GPT-4生成所有候选答案，对生成的结果进行N选1，最终根据答案来评分

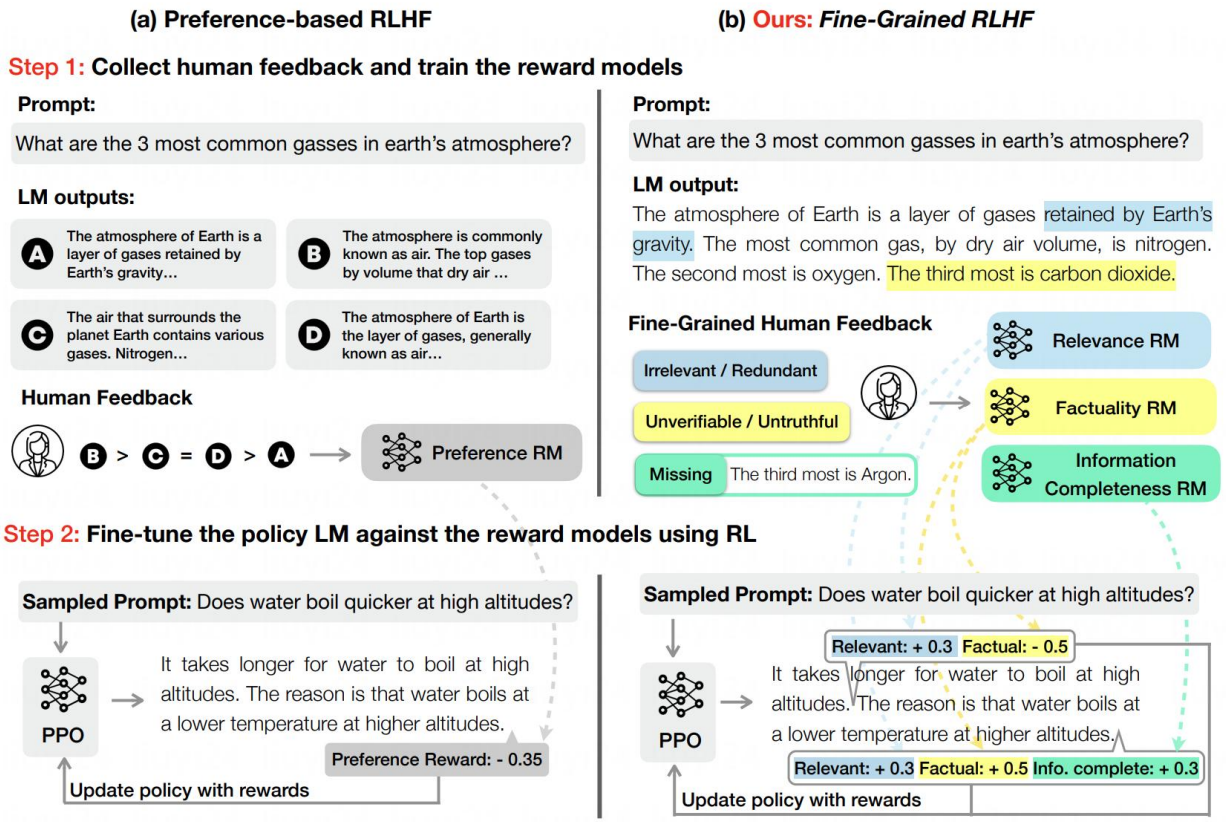
Q：我理解工业界的工作有需要遵守的保密规定，但作者们或许可以在他们的预训练和微调数据集上进行开源模型的实验？由于可重现性问题，我认为这项工作更适合作为blog而不是学术论文

A：我们的大规模结果无法在当前任何可用的开源模型上复制。这是当前技术状态的一个不幸限制，也是一个有效的批评。我们仍然相信，汇报大规模的前沿结果是很重要的。这表明了在所有的模型规模下，过程监督都能使奖励模型的训练更具成本效益

Let's Verify Step by Step. CloseAI. ICLR 2024 under review

3.2 多目标的人类偏好对齐

- 单一的奖励值无法准确反映多样化的人类偏好与价值观，应当在对齐时进行解耦，使用不同的Reward Model来表示不同的人类偏好



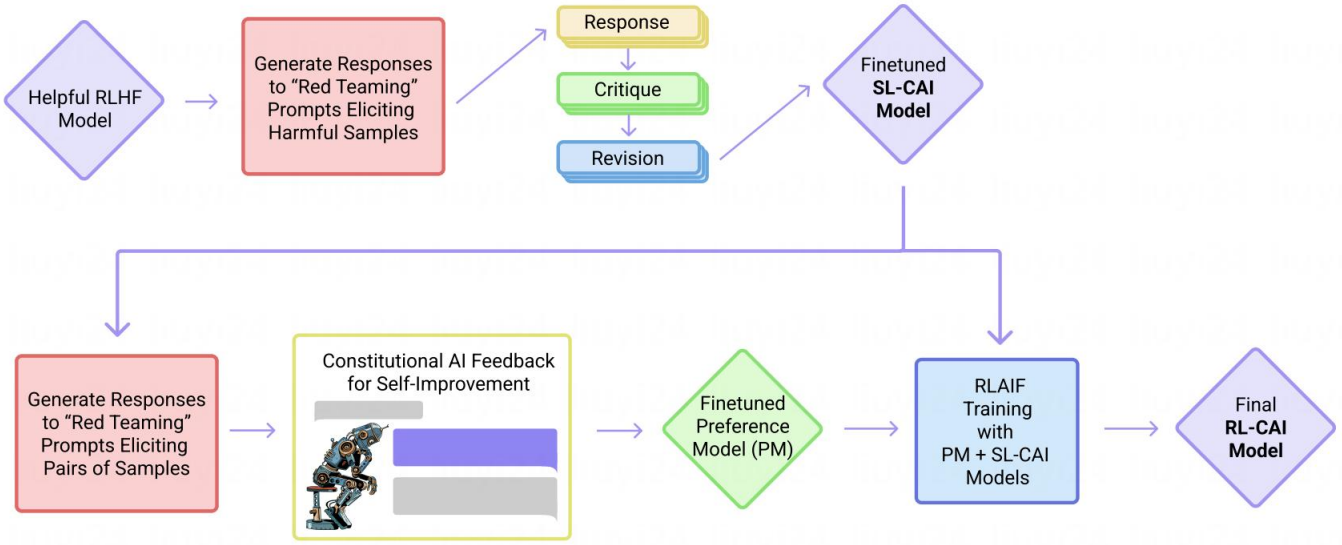
- 不相关、重复或不连贯
- 逻辑或事实性错误
- 全文信息不完整

Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. Allen AI. NIPS 2023 poster

3.3 自动对齐

- 相较于人类反馈，目前的通用AI系统表现出了更好的创造力与可拓展性
- 将进行对齐工作所需的任务，逐渐从人类转交给一个自动化的AI系统，使得Researcher能够将更多的时间和精力，投入到更具挑战性的领域

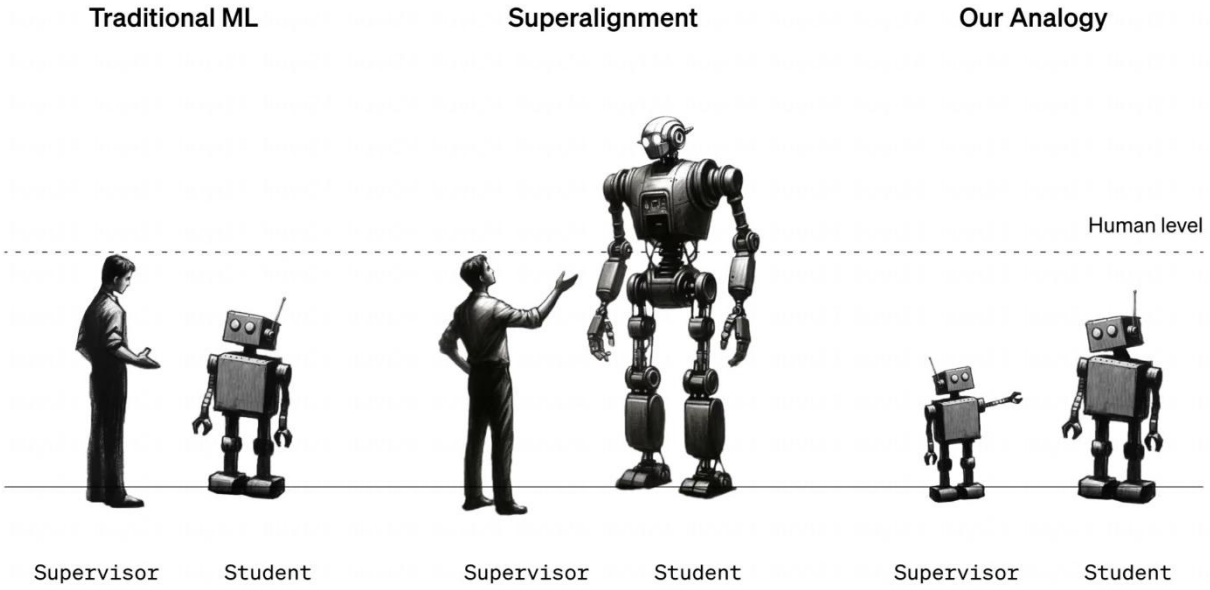
- **宪法AI：**使用LLM替代人工标注，以监督人类偏好对齐的训练流程
  - Phase1: Harmful Resp from Helpful RLHF → Critique → Revision → SL-**CAI Model**
  - Phase2: Harmless from CAI + Helpful from Human → RM → RL-**CAI Model**



Constitutional AI: Harmlessness from AI Feedback. Anthropic. research paper

3.4 如何监督远超人类的大模型

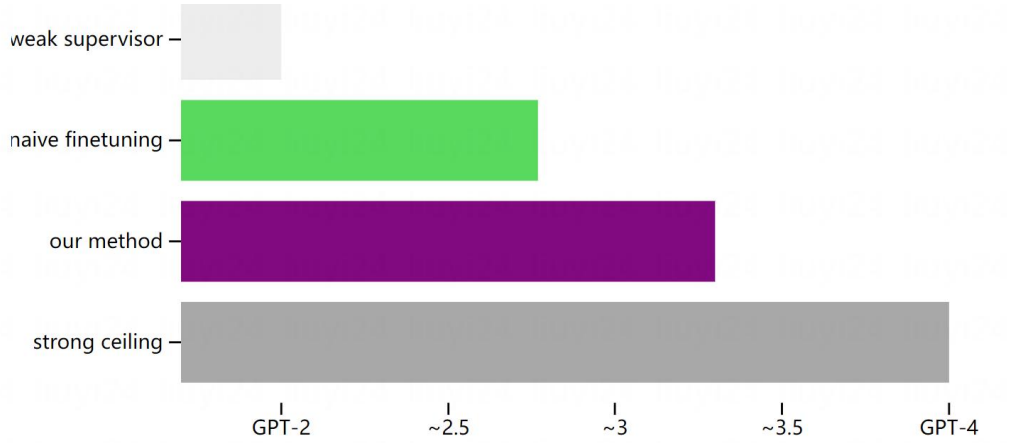
- 此前的对齐方法依赖于人类监督，但随着大模型的不断发展，或许能够做出极其复杂和富有创造性的行为，使人类难以对其进行可靠的监督
- OpenAI的超级对齐小组（superalignment）通过研究一个类似的问题：使用小模型监督大模型，来对人类是否能有效地监督超级人工智能的原问题进行实证研究



Weak-to-strong generalization: Eliciting Strong Capabilities with Weak Supervision. OpenAI. research paper

3.4 如何监督远超人类的大模型

- 实验方法：
  - **弱监督模型**-基于标注数据训练得到的小模型，扮演人类角色
  - **强学生模型**-基于弱监督者生成的合成数据训练得到的大模型
  - **上限模型**-基于标注数据训练得到的大模型



- 研究表明，小模型如GPT-2可以被用来激发GPT-4的大部分能力，使其达到接近GPT-3.5的性能，甚至可以正确地泛化到小模型失败的难题上
- OpenAI将这种现象称为弱到强泛化（Weak-to-strong generalization），这表明大模型具备如何执行任务的隐含知识，即使在给出低质量甚至错误的监督信号时

Weak-to-strong generalization: Eliciting Strong Capabilities with Weak Supervision. OpenAI. research paper

谢谢大家！

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS