REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

基于人类反馈的强化学习

报告人： 毕冠群/时间：2023/3/3

目录

目录

ChatGPT的发展历程

- GPT-1 用的是无监督预训练 + 有监督微调。
- GPT-2 用的是纯无监督预训练。
- GPT-3 沿用了 GPT-2 的纯无监督预训练，但是数据大了好几个量级。
- InstructGPT 在 GPT-3 上用强化学习做微调，内核模型为 PPO-ptx。
- ChatGPT 沿用了 InstructGPT，但是数据大了好几个量级，并试图减少有害和误导性的回复

ChatGPT训练过程

Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

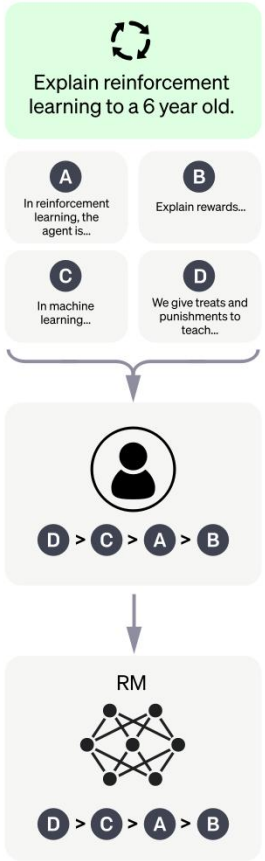
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

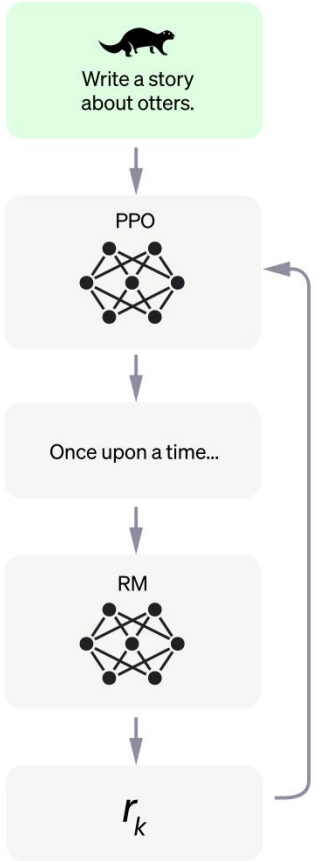
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

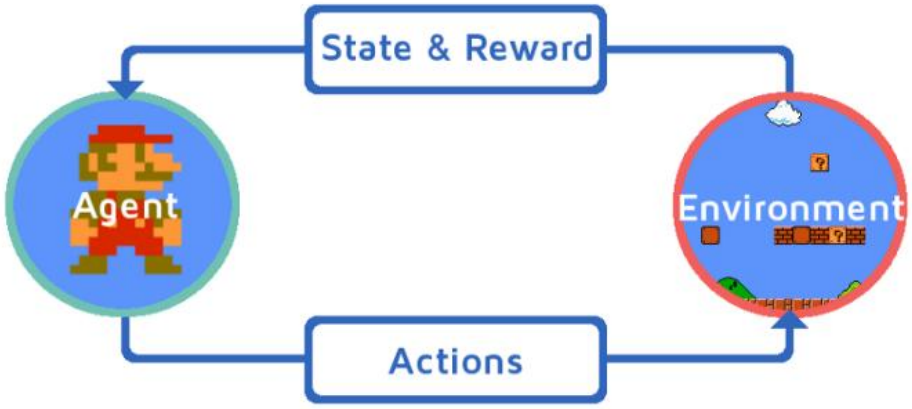


https://openai.com/blog/chatgpt/

目录

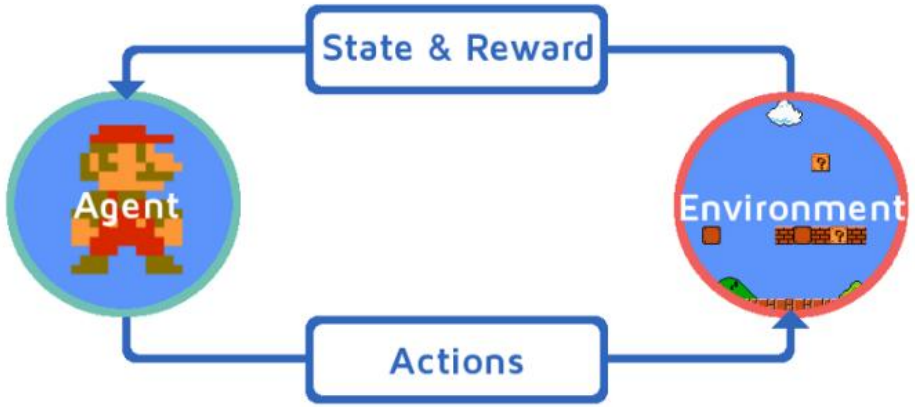
强化学习

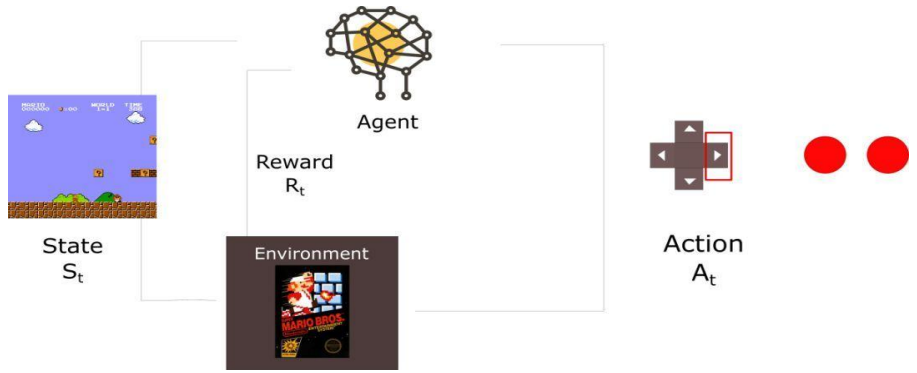
- 强化学习（Reinforcement Learning, RL）是机器学习的范式和方法论之一，用于描述和解决智能体在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题 。
- **中心思想**:让智能体在环境里学习。每个行动会对应各自的奖励，智能体通过分析数据来学习，怎样的情况下应该做怎样的事情。
- **序列决策问题**：一个决策代理（decision agent）与离散的时间动态系统进行迭代地交互。 在每个时间步的开始时，系统会处于某种状态。基于代理的决策规则，它会观察当前的状态，并从有限状态集中选择一个。然后，动态系统会进入下一个新的状态并获得一个对应的收益。这样循环进行状态选择，以获得一组最大化收益。



强化学习

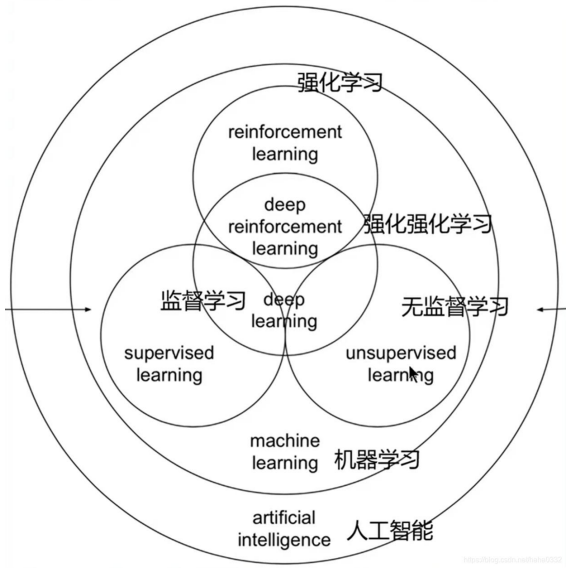
- **基本要素**
  1. **状态**(State)：是对环境的描述，可以是离散的或连续的
  2. **行为**（Action）：环境中智能体某一轮采取的动作，这个动作会作用于环境，且执行后将获得一个奖励（可正可负）。
  3. **奖励**（Reward）：本质就是为了完成某一目标的动作质量。从长远的角度看什么是好的，一个状态的价值是一个智能体从这个状态开始，对将来累积的总收益的期望。
  4. **策略**（Policy）：策略定义了智能体在特定时间的行为方式，即，策略是环境状态到动作的映射。
  5. **回合**(Episode)：智能体在环境里面执行某个策略从开始到结束这一过程。
- **两个可以进行交互的对象**
  1. **智能体**（Agent）：感知环境状态（State），根据反馈奖励（Reward）选择合适行为（Action）最大化长期收益，在交互过程中进行学习
  2. **环境**（Environment）：游戏发生的场景，可以被智能体做出的动作改变。接收智能体执行的一系列动作，对这一系列动作进行评价并转换为一种可量化的信号，最终反馈给智能体。环境中有一个一个的状态(State)

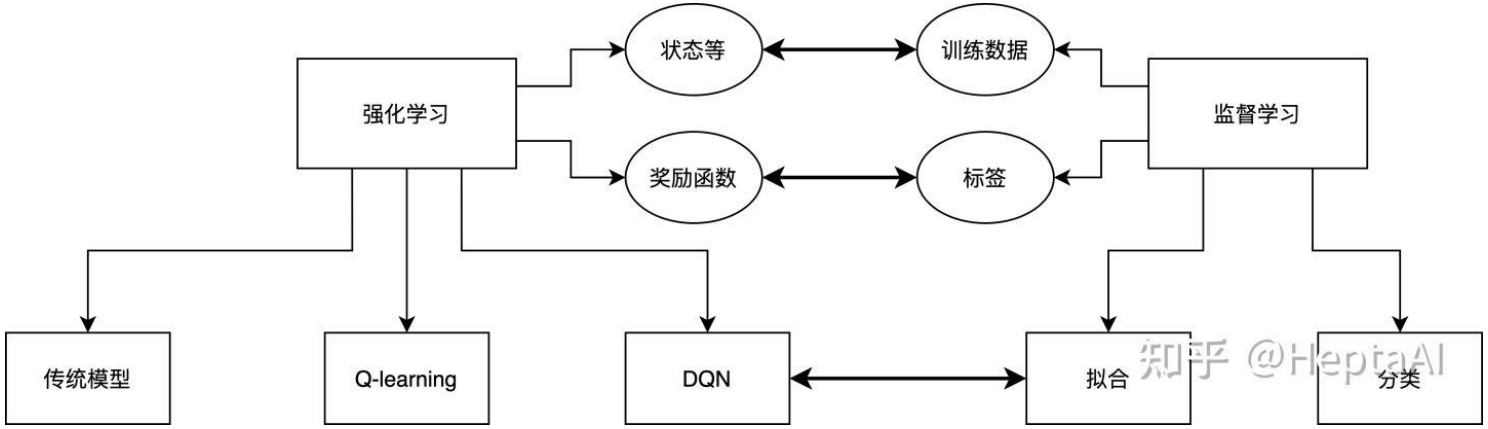




强化学习的特点

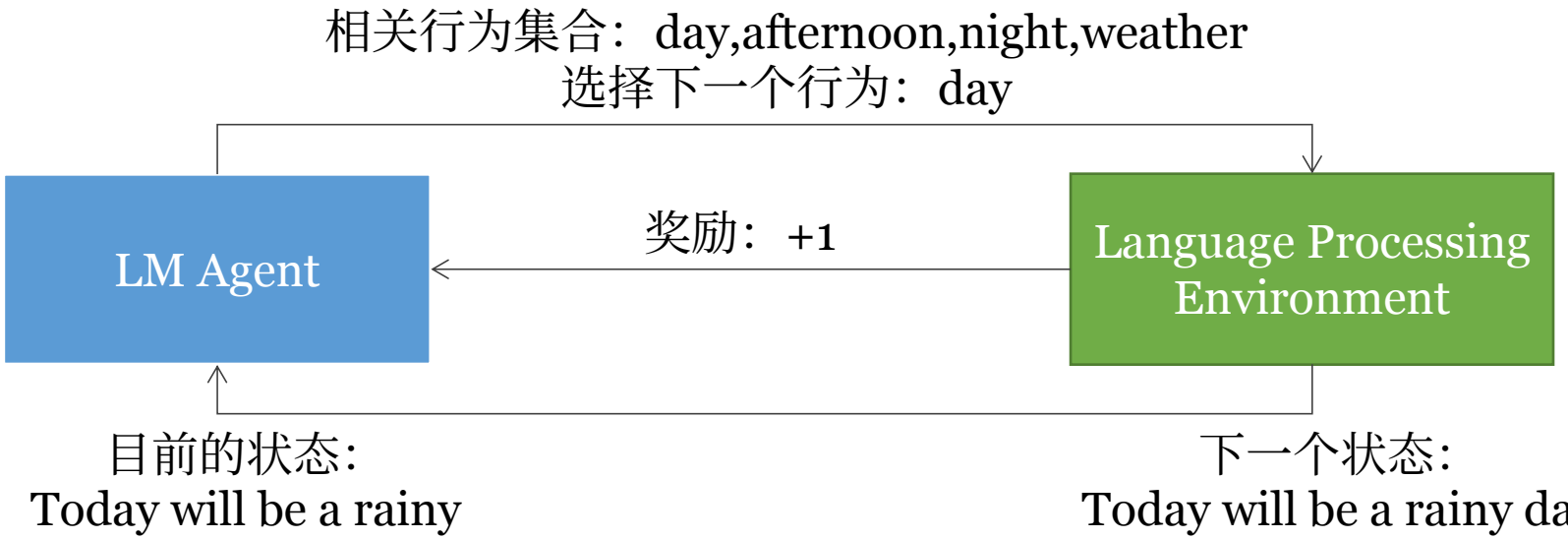
- 监督学习：从外部监督者提供的带标注训练集中进行学习（任务驱动）
- 无监督学习：寻找未标注数据中隐含结构（数据驱动）
- 强化学习：探索-开发权衡 (从错误中学习)
  1. 处理序列数据，不满足独立同分布
  2. 没有对错，只有奖励
  3. 延迟奖励，追求序列最优，而不是任意一次行为都要最优
  4. 需要试错探索
  5. 数据分布随模型更新而产生变化





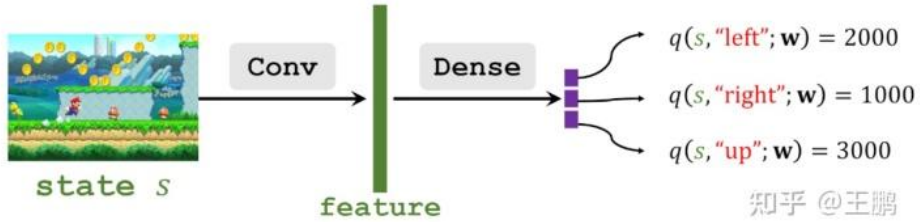
强化学习+文本生成

- 文本生成可以建模为一个token空间上的序列决策问题（选择一个token后继续选择另一个token）
  - State：对话上下文
  - Action：回复的token space上的token
  - Reward：生成的质量判别
  - Episode：一次完整的解码生成回复的过程

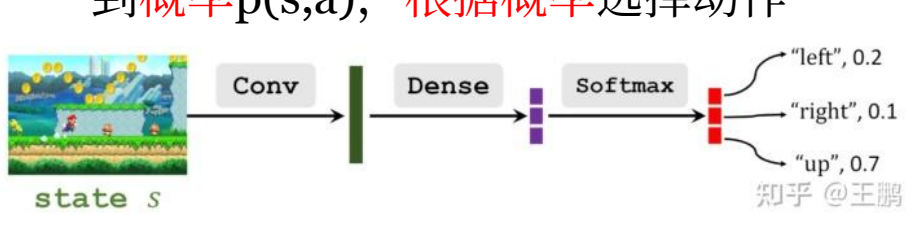


强化学习分类

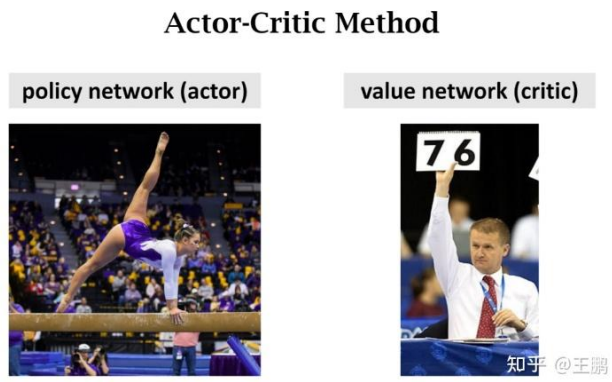
1. 基于值，Value-based
- 计算动作期待值，选取期待值最大的动作



2. 基于策略，Policy Gradient
- 有一个函数计算此刻选择哪个动作，并得到概率p(s,a)，根据概率选择动作



3. Actor-Critic 融合了上述两种方法，价值函数和策略函数一起进行优化。
- 价值函数负责在环境学习并提升自己的价值判断能力
- 策略函数则接受价值函数的评价，尽量采取在价值函数那可以得到高分的策略。

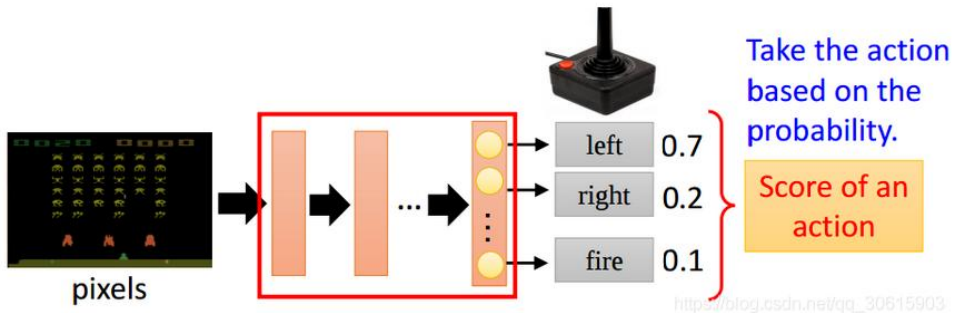


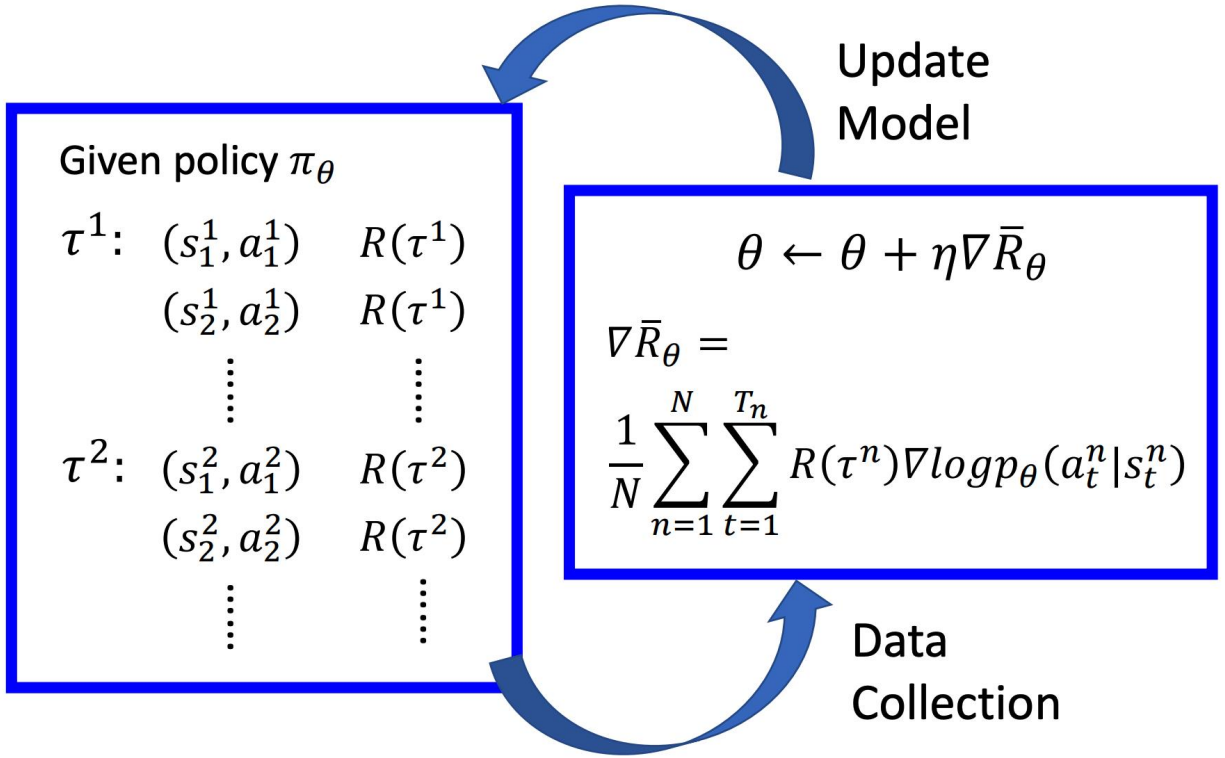
目录

Policy Gradient

目的：直接建模与优化policy。

过程：
Actor和Environment进行互动，产生一系列采样数据，即获得很多（s，a）对（表示在状态s下采取动作a，得到当前奖励R(τ)），然后将这些数据送入训练过程中计算，并更新模型的参数θ，如此循环往复。





PPO 近端策略优化

改进原因

- 更新策略需要用到从当前策略中采样的最新的样本，因此每次只更新一步便把样本舍弃，这意味着旧样本不能重复使用，训练效率低。
- PPO算法利用重要性采样的思想，在不知道策略路径的概率p的情况下，通过模拟一个近似的q分布，只要p同q分布不差的太远，通过多轮迭代可以快速参数收敛

importance sampling

$$
\begin{aligned}
E_{x\sim p(x)}[f(x)] &= \int p(x)f(x)dx \\
&= \int \frac{q(x)}{q(x)}p(x)f(x)dx \\
&= \int q(x)\frac{p(x)}{q(x)}f(x)dx \\
&= E_{x\sim q(x)}\left[\frac{p(x)}{q(x)}f(x)\right]
\end{aligned}
$$

Algorithm 1 PPO, Actor-Critic Style

for iteration=$1, 2, \ldots$ **do**
  for actor=$1, 2, \ldots, N$ **do**
    Run policy $\pi_{\theta_{\text{old}}}$ in environment for T timesteps
    Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
  end for
  Optimize surrogate L wrt θ, with K epochs and minibatch size $M \leq NT$
  $\theta_{\text{old}} \leftarrow \theta$
end for

重要性采样&存在的问题

$$E_{x\sim p}[f(x)] = E_{x\sim q}[f(x)\frac{p(x)}{q(x)}]$$

$$\mathrm{Var}_{x\sim p}[f(x)] \quad \mathrm{Var}_{x\sim q}[f(x)\frac{p(x)}{q(x)}]$$

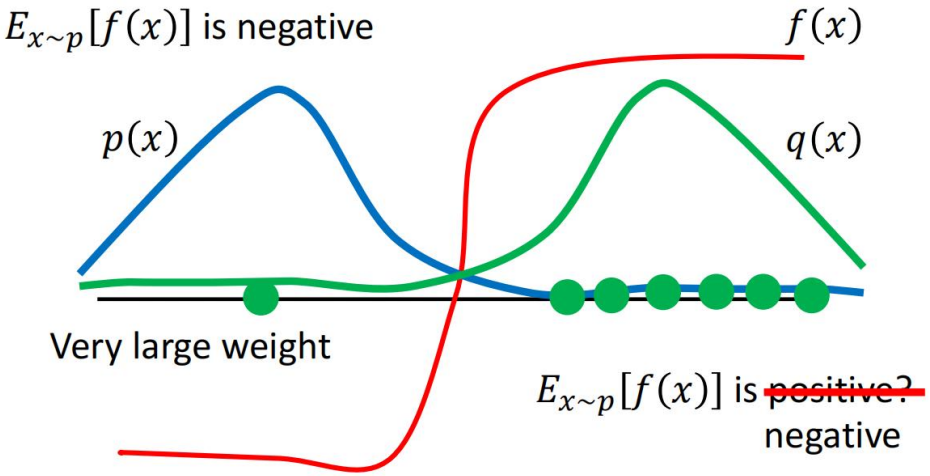
$$VAR[X] = E[X^2] - (E[X])^2$$

$$\mathrm{Var}_{x\sim p}[f(x)] = E_{x\sim p}[f(x)^2] - \left(E_{x\sim p}[f(x)]\right)^2$$

$$\mathrm{Var}_{x\sim q}[f(x)\frac{p(x)}{q(x)}] = E_{x\sim q}\left[\left(f(x)\frac{p(x)}{q(x)}\right)^2\right] - \left(E_{x\sim q}\left[f(x)\frac{p(x)}{q(x)}\right]\right)^2$$

$$= E_{x\sim p}\left[f(x)^2\frac{p(x)}{q(x)}\right] - \left(E_{x\sim p}[f(x)]\right)^2$$

$$E_{x\sim p}[f(x)] = E_{x\sim q}\left[f(x)\frac{p(x)}{q(x)}\right]$$



重要性采样应用于PG

$$\nabla \bar{R}_\theta = E_{\tau \sim p_\theta(\tau)}[R(\tau)\nabla log p_\theta(\tau)] \quad \nabla \bar{R}_\theta = E_{\tau \sim p_{\theta'}(\tau)}\left[\frac{p_\theta(\tau)}{p_{\theta'}(\tau)} R(\tau)\nabla log p_\theta(\tau)\right]$$

$$\nabla f(x) = f(x)\nabla log f(x)$$

Gradient for update

$$= E_{(s_t,a_t)\sim\pi_\theta}[A^\theta(s_t,a_t)\nabla log p_\theta(a_t^n|s_t^n)]$$

$A^{\theta'}(s_t,a_t)$ This term is from sampled data.

$$= E_{(s_t,a_t)\sim\pi_{\theta'}}[\frac{P_\theta(s_t,a_t)}{P_{\theta'}(s_t,a_t)} A^\theta(s_t,a_t)\nabla log p_\theta(a_t^n|s_t^n)]$$

$$= E_{(s_t,a_t)\sim\pi_{\theta'}}[\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)} \frac{p_\theta(s_t)}{p_{\theta'}(s_t)} A^\theta(s_t,a_t)\nabla log p_\theta(a_t^n|s_t^n)]$$

$$J^{\theta'}(\theta) = E_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)} A^{\theta'}(s_t,a_t)\right]$$

PPO algorithm

$$J^{\theta^k}(\theta) \approx \sum_{(s_t,a_t)} \frac{p_\theta(a_t|s_t)}{p_{\theta^k}(a_t|s_t)} A^{\theta^k}(s_t,a_t)$$

- Initial policy parameters θ^0
- In each iteration
  - Using θ^k to interact with the environment to collect $\{s_t, a_t\}$ and compute advantage $A^{\theta^k}(s_t,a_t)$
  - Find θ optimizing $J_{PPO}(\theta)$

$$J_{PPO}^{\theta^k}(\theta) = J^{\theta^k}(\theta) - \beta KL(\theta,\theta^k)$$

Update parameters several times

- If $KL(\theta,\theta^k) > KL_{max}$, increase β
- If $KL(\theta,\theta^k) < KL_{min}$, decrease β

Adaptive KL Penalty

PPO-Background

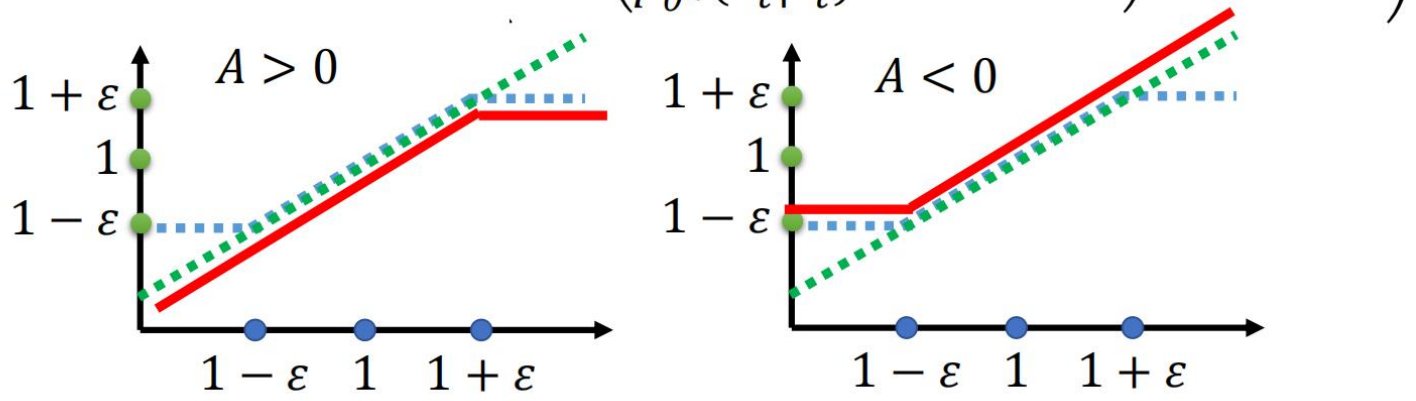
PPO algorithm

$$J_{PPO}^{\theta^k}(\theta) = J^{\theta^k}(\theta) - \beta KL(\theta, \theta^k)$$

$$J^{\theta^k}(\theta) \approx \sum_{(s_t, a_t)} \frac{p_\theta(a_t|s_t)}{p_{\theta^k}(a_t|s_t)} A^{\theta^k}(s_t, a_t)$$

PPO2 algorithm

$$J_{PPO2}^{\theta^k}(\theta) \approx \sum_{(s_t, a_t)} min\left(\frac{p_\theta(a_t|s_t)}{p_{\theta^k}(a_t|s_t)} A^{\theta^k}(s_t, a_t), clip\left(\frac{p_\theta(a_t|s_t)}{p_{\theta^k}(a_t|s_t)}, 1-\varepsilon, 1+\varepsilon \right) A^{\theta^k}(s_t, a_t) \right)$$



PPO-Experiment Result

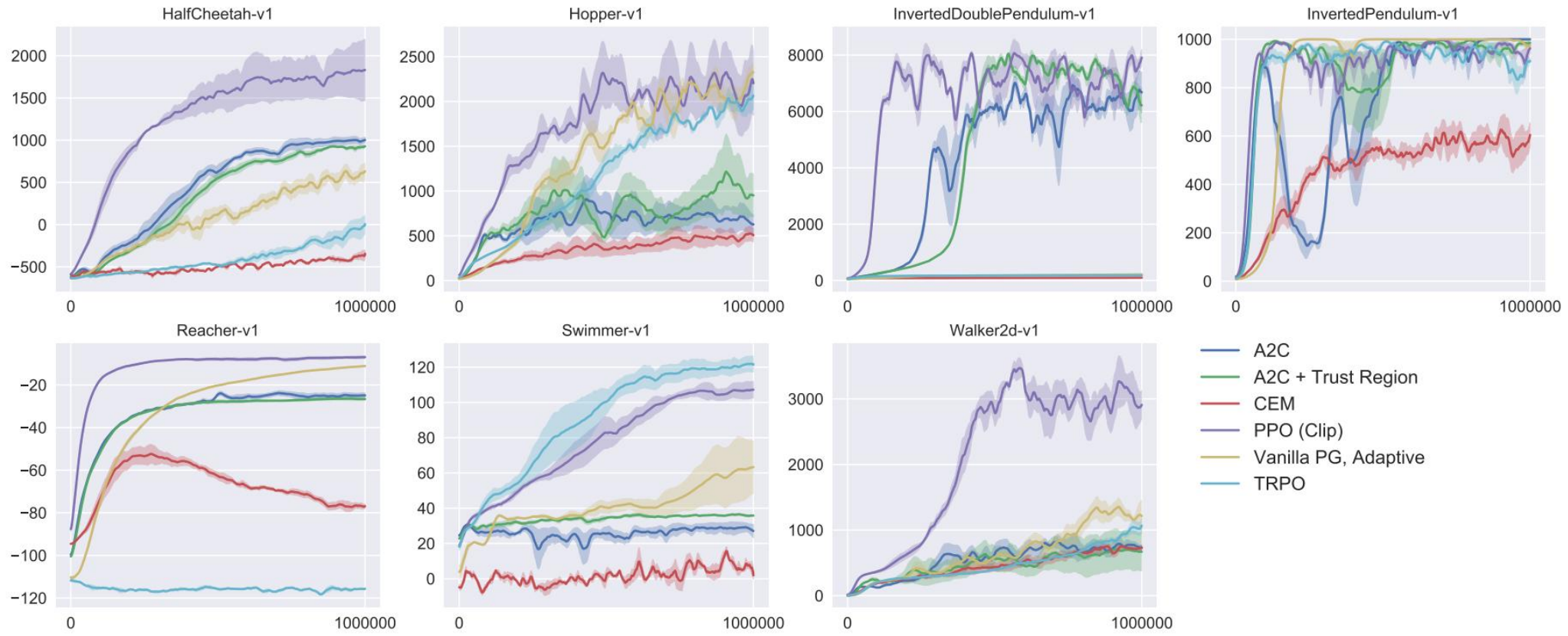
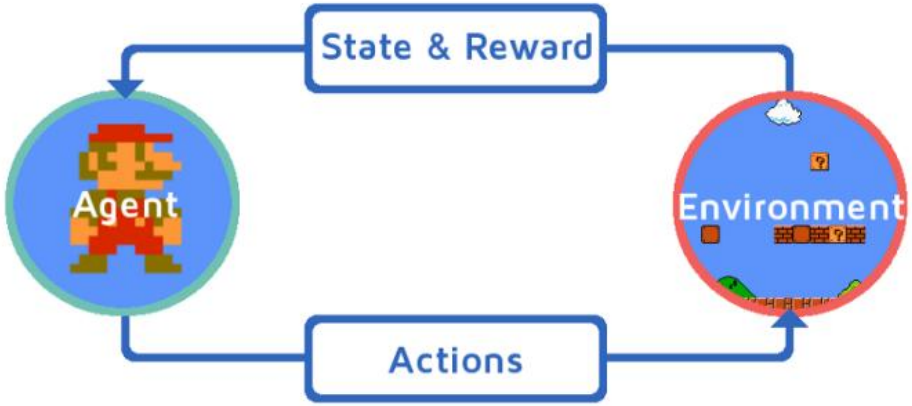
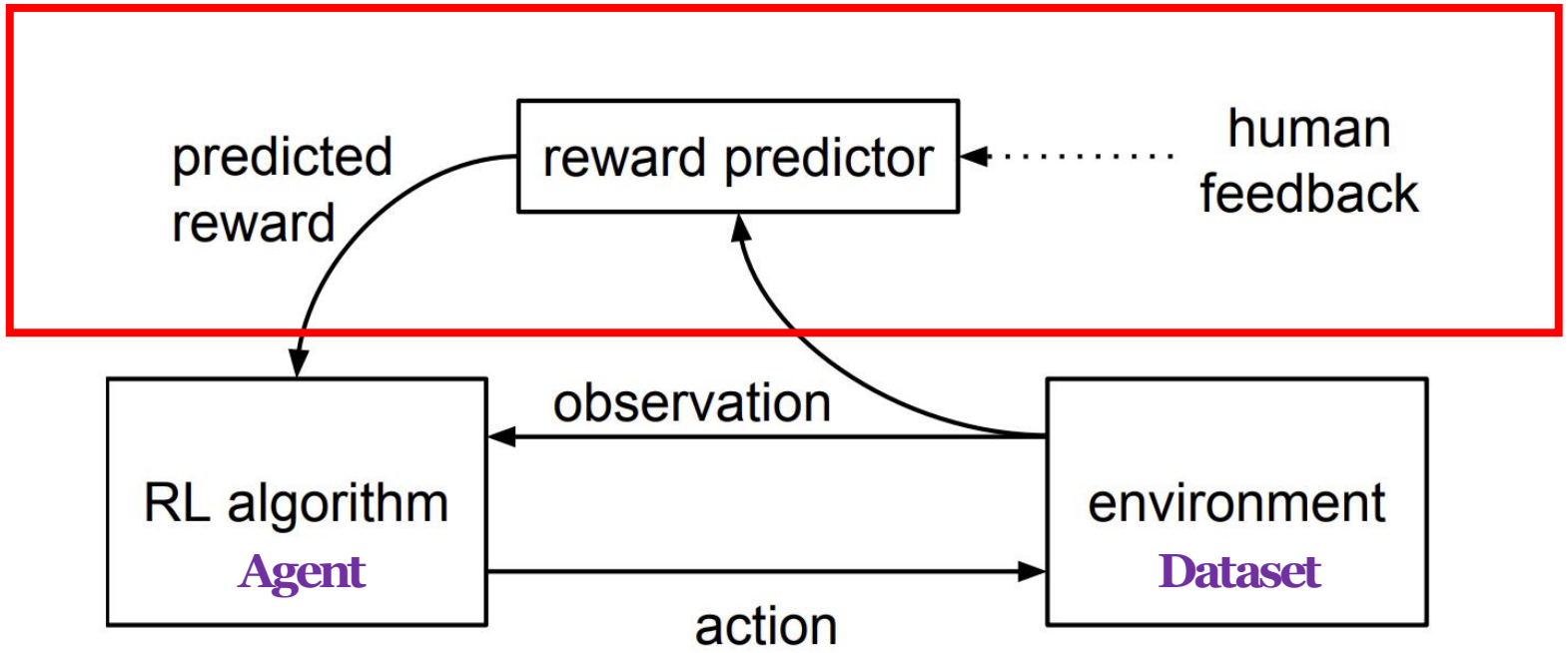


Figure 3: Comparison of several algorithms on several MuJoCo environments, training for one million timesteps.

目录

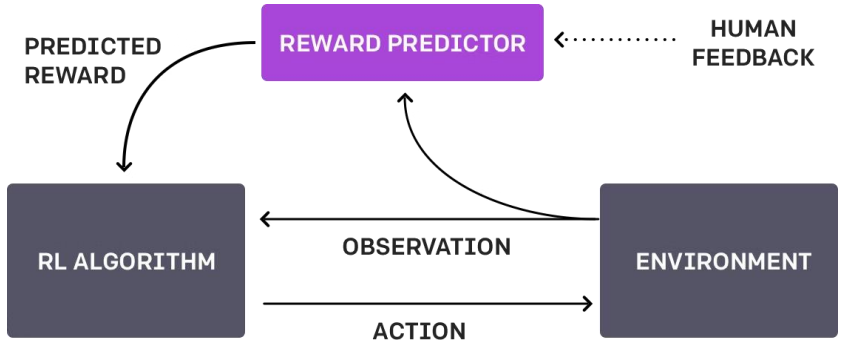
Naive RL & RLHF



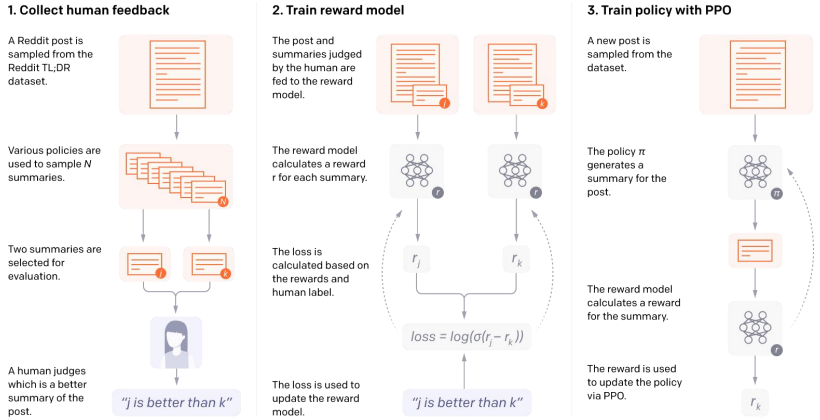


RLHF发展历程

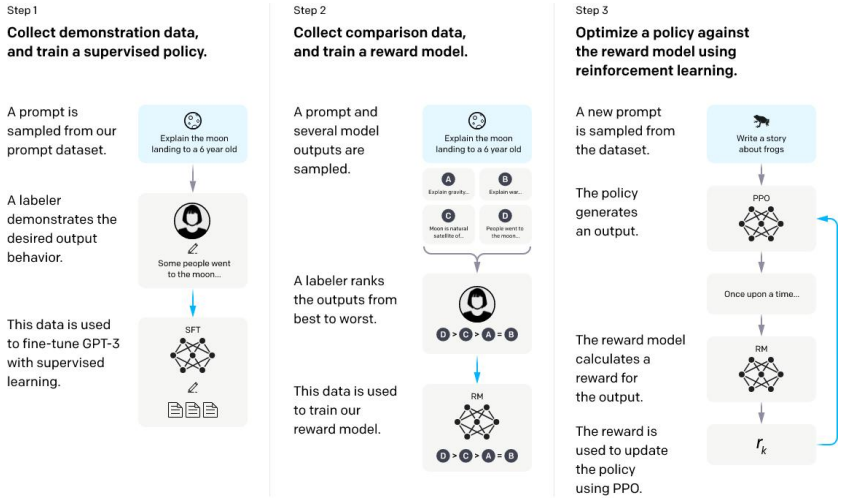
1. NIPS 2017 机器人&Atari



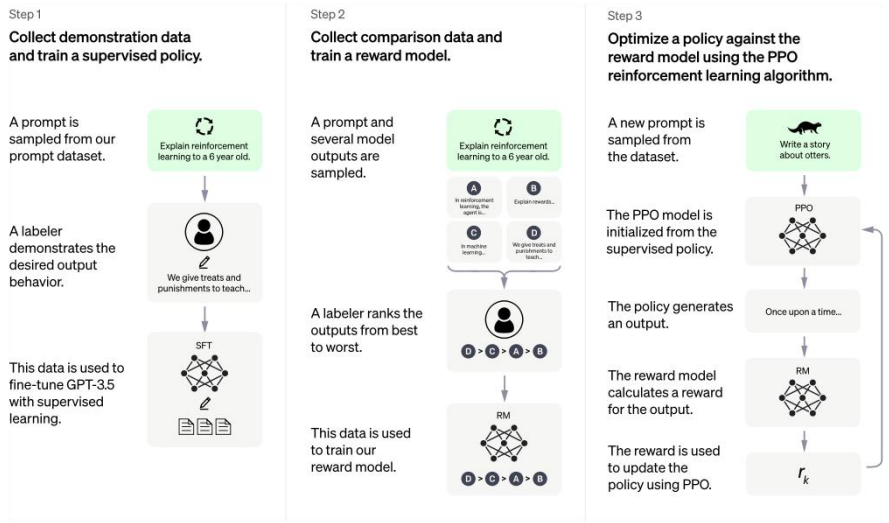
2. NIPS 2020 文本摘要



3. 2022.3 InstructGPT



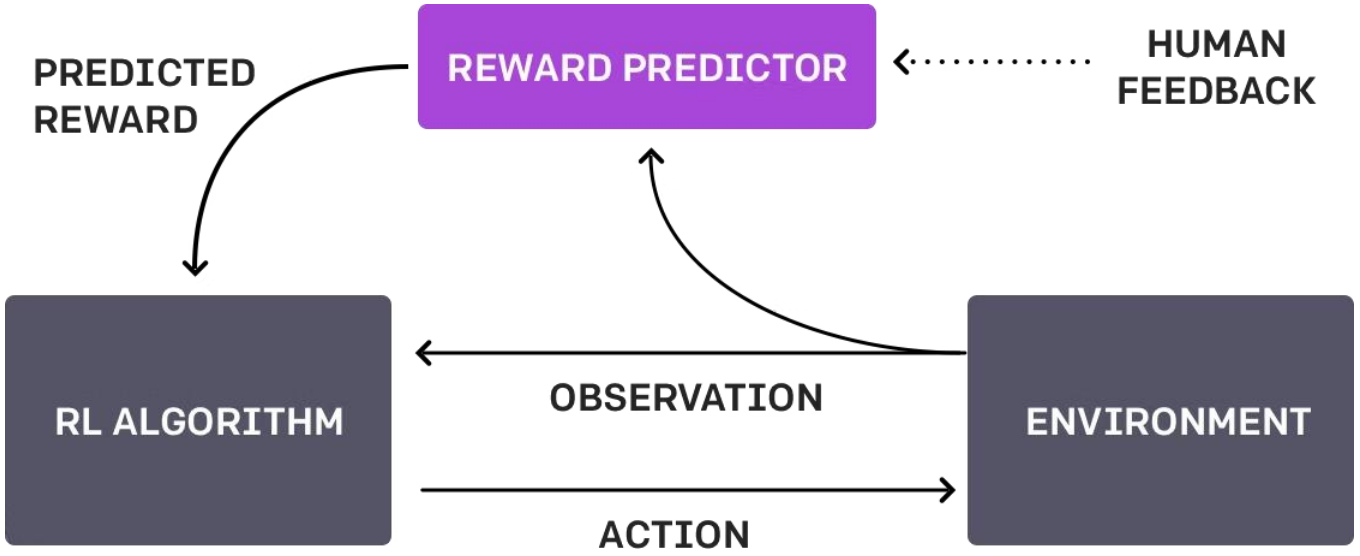
4. 2022.11 ChatGPT



DRL from Human Preferences, NIPS 2017, DeepMind & OpenAI

现存问题：

- 许多任务涉及复杂、定义不明确或难以指定的目标
- 不合适的奖励函数常常会导致学到的行为实际上不满足我们的偏好。 **This difficulty underlies recent concerns about misalignment between our values and the objectives of our RL systems**
- 逆强化学习不直接适用于人类难以执行的行为
- 虽然我们不能给出复杂任务的完整轨迹，但是可以评估某条轨迹的效果；
  给出稳定的定量评估分数难，定性地对比两条轨迹哪个比较好容易



主要贡献：

将人类反馈扩展到深度强化学习，并学习更复杂的行为

应用领域：

- Arcade 学习环境中的 Atari 游戏
- 物理模拟器 MuJoCo 中的机器人任务

DRL from Human Preferences, NIPS 2017, DeepMind & OpenAI

设置与目标

轨迹片段是观察和动作组成的序列 $\sigma = ((o_0, a_0), (o_1, a_1), \ldots, (o_{k-1}, a_{k-1})) \in (\mathcal{O} \times \mathcal{A})^k$.

$\sigma^1 \succ \sigma^2$表示相比轨迹 σ^1更喜欢轨迹 σ^2

具体方法

维护一个策略$\pi : \mathcal{O} \to \mathcal{A}$ 和一个奖励函数 $\hat{r} : \mathcal{O} \times \mathcal{A} \to \mathbb{R}$，二者都由DNN参数化

1. 通过策略π与环境相互作用产生一组轨迹{τ^1,τ^2,...,τ^i}.π的参数通过传统的 RL 算法进行更新，以最大化预测奖励 $r_t = \hat{r}(o_t, a_t)$
2. 从第一步产生的轨迹{τ^1,τ^2,...,τ^i}中选择成对的轨迹片段(σ^1,σ^2),，并将其发送给人类进行比较
3. 所有历史片段比较样本放在一起，用监督学习得到映射 $\hat{r}$

优化策略：TRPO

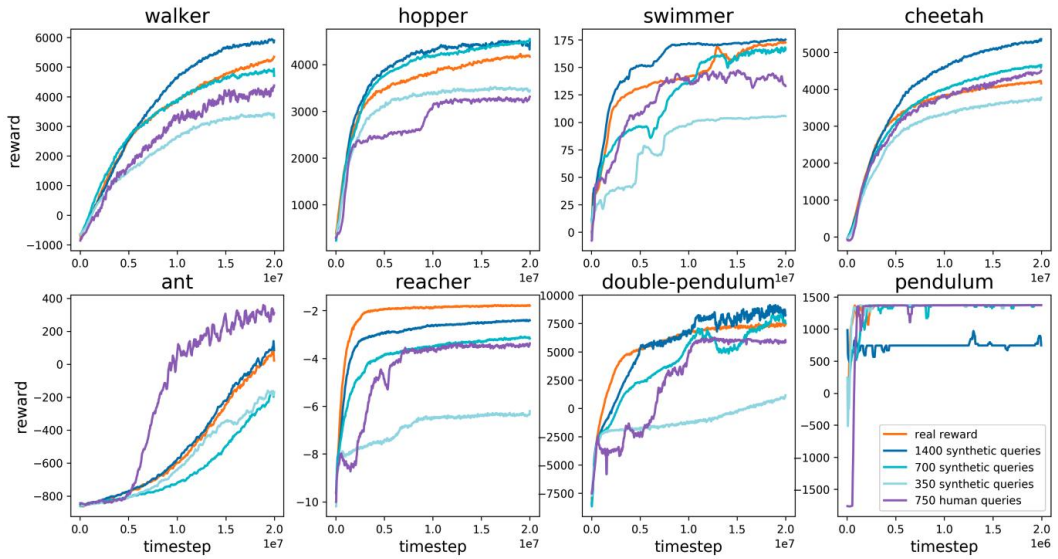
偏好诱导：观看两段轨迹，指出更喜欢哪一个/同样好/无法比较。$(\sigma^1, \sigma^2, \mu)$

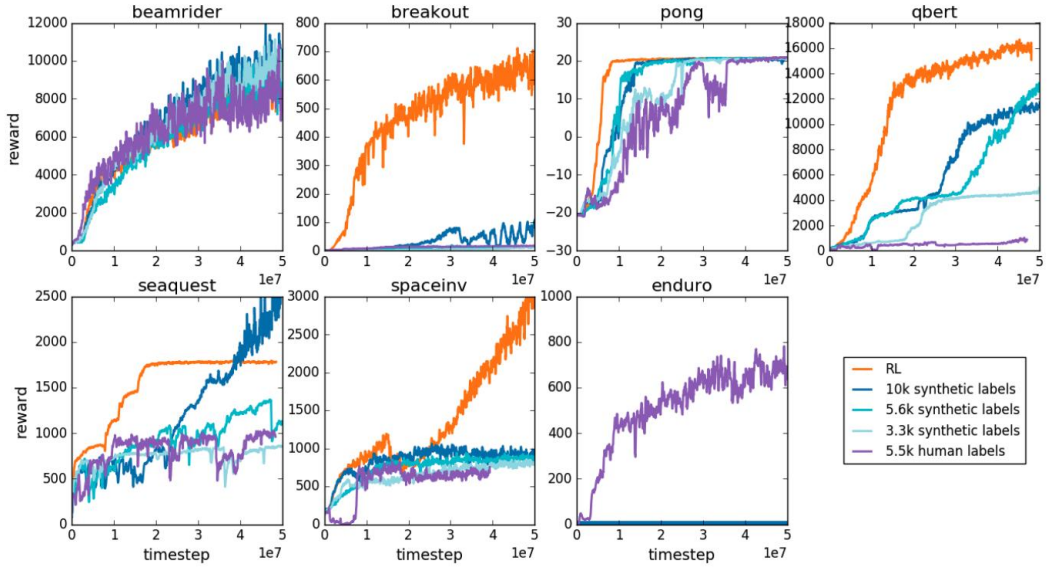
拟合奖励函数:最小化预测偏好分布和实际人类偏好分布之间的交

$$\hat{P}\left[\sigma^1 \succ \sigma^2\right] = \frac{\exp \sum \hat{r}\left(o_t^1, a_t^1\right)}{\exp \sum \hat{r}(o_t^1, a_t^1) + \exp \sum \hat{r}(o_t^2, a_t^2)}.$$

$$\operatorname{loss}(\hat{r}) = -\sum_{(\sigma^1, \sigma^2, \mu) \in \mathcal{D}} \mu(1) \log \hat{P}\left[\sigma^1 \succ \sigma^2\right] + \mu(2) \log \hat{P}\left[\sigma^2 \succ \sigma^1\right].$$

DRL from Human Preferences, NIPS 2017, DeepMind & OpenAI





blog:https://openai.com/research/learning-from-human-preferences

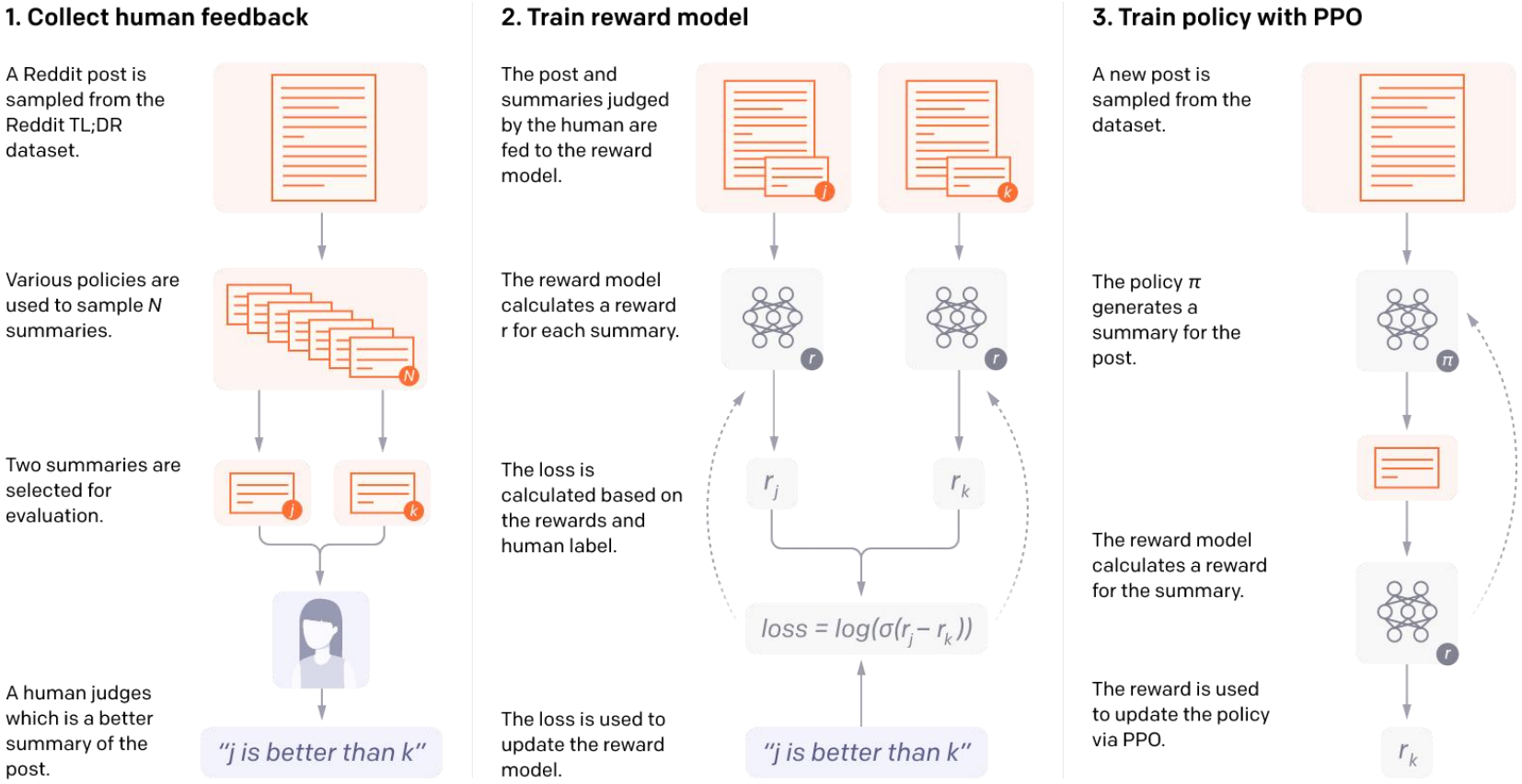
Summary with HF, OpenAI, NIPS 2020

问题

- 自动指标难以刻画对摘要质量的要求；
- 难以对较为复杂的NLP任务进行评测、构造精准的损失函数

方法

- 以人类偏好替代自动化评测方法（如ROUGE、BLEU）为训练目标，用人类反馈作为奖励进行强化学习



Learning to summarize from human feedback, NIPS 2020

Summary with HF, NIPS 2020

数据

摘要数据:Reddit, TL;DR数据集
（没有使用更简单的CNN/DM）

任务

- 摘要长度≤48；
- 忠实

收集人类反馈

- 完全过渡到离线设置
- 与标注员保持密切联系
- 训练标注员，提升与研究者的一致性
- 标注人员来自众包平台(freelancing platform)

模型

Backbone使用GPT-3 1.3B, 6.7B

- Pretrained models(zero-shot baselines)
- Supervised baselines

What makes for a good summary? Roughly speaking, a good summary is a shorter piece of text that has the essence of the original – tries to accomplish the same purpose and conveys the same information as the original post. We would like you to consider these different dimensions of summaries:

Essence: is the summary a good representation of the post?

Clarity: is the summary reader-friendly? Does it express ideas clearly?

Accuracy: does the summary contain the same information as the longer post?

Purpose: does the summary serve the same purpose as the original post?

Concise: is the summary short and to-the-point?

Style: is the summary written in the same style as the original post?

Generally speaking, we give higher weight to the dimensions at the top of the list. Things are complicated though – none of these dimensions are simple yes/no matters, and there aren't hard and fast rules for trading off different dimensions. This is something you'll pick up through practice and feedback on our website.

Table 6: An excerpt from the instructions we gave to labelers for doing comparisons.

Summary with HF, NIPS 2020

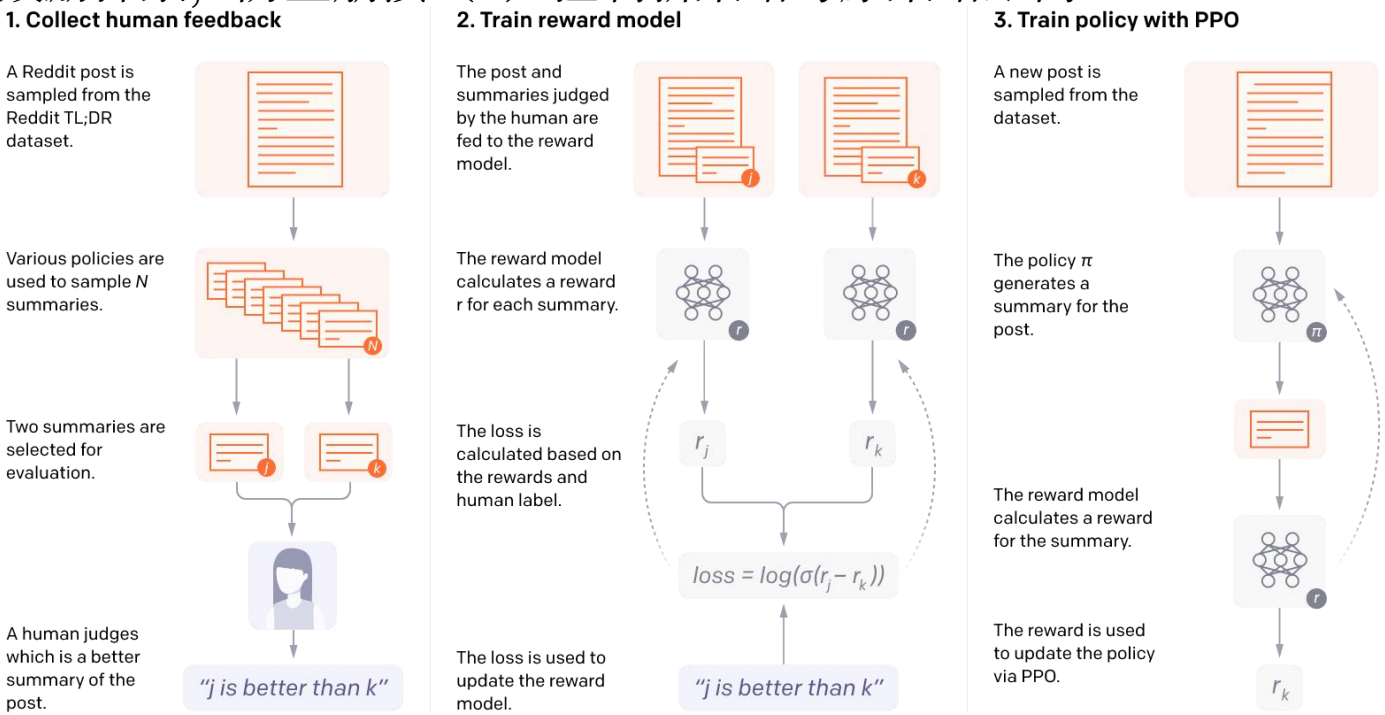
1. LM & Data

2. Reward Model $\text{loss}(r_\theta) = -E_{(x,y_0,y_1,i)\sim D}[\log(\sigma(r_\theta(x,y_i) - r_\theta(x,y_{1-i})))]$

- supervised baseline+随机初始化linear head；
- 输出标量reward

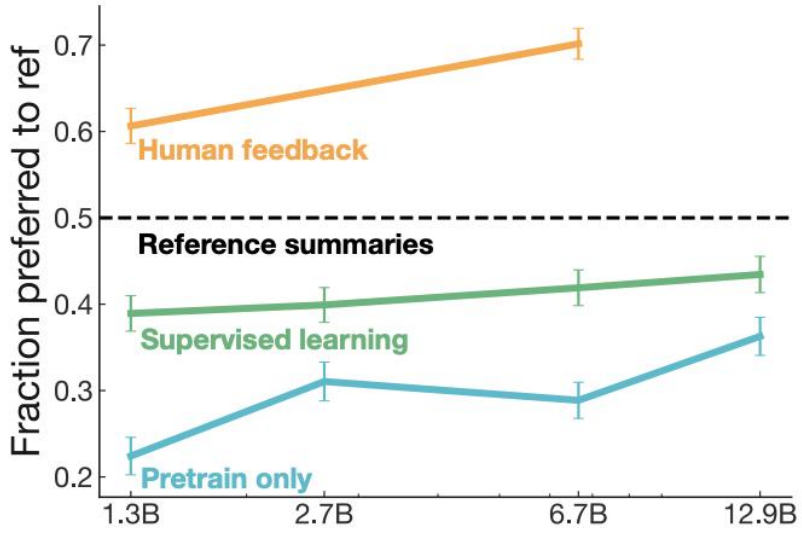
3. Human Feedback Policies $R(x,y) = r_\theta(x,y) - \beta\log[\pi_\phi^{\text{RL}}(y|x)/\pi^{\text{SFT}}(y|x)]$

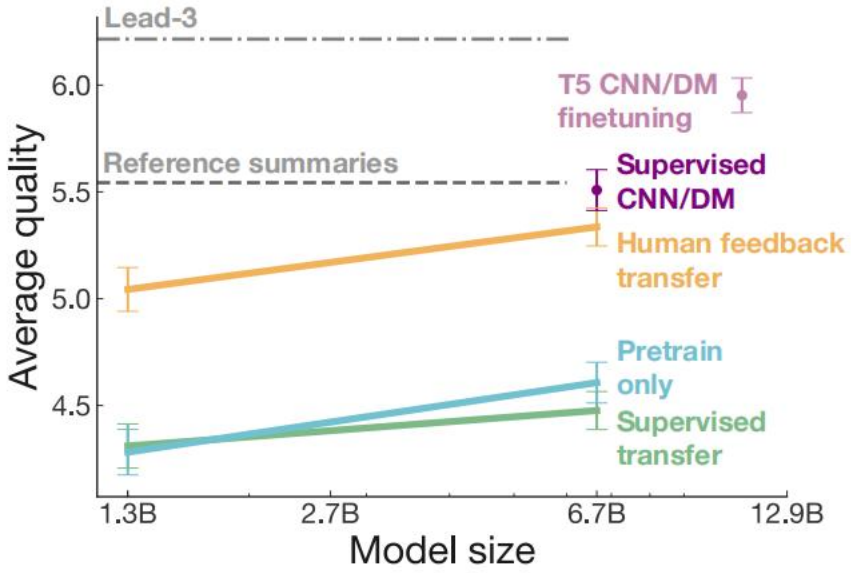
KL散度项（1）鼓励探索，防止崩溃（2）控制新策略与原策略距离

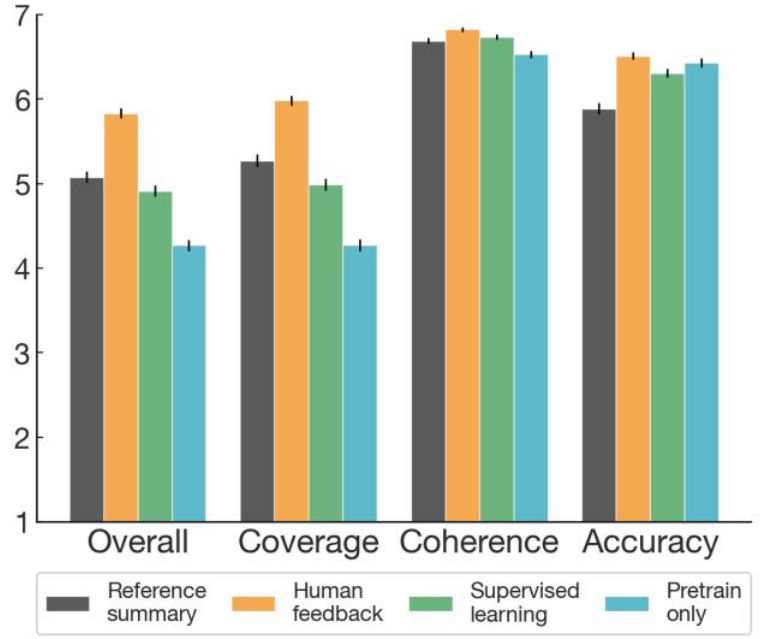


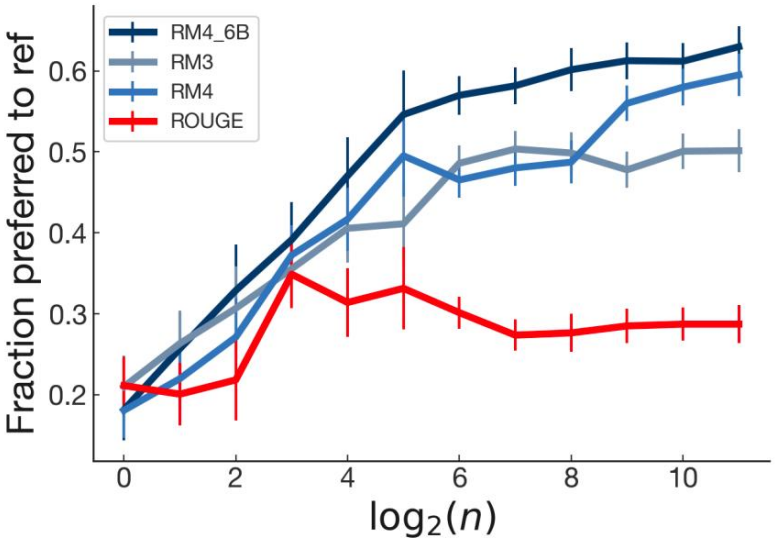
Learning to summarize from human feedback, NIPS 2020

RLHF for Summarization









目录

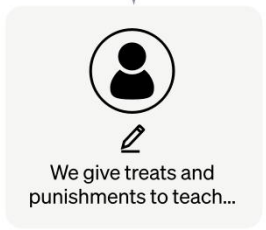
Step 1

Step 1

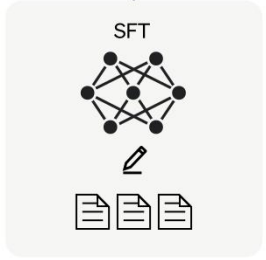
Collect demonstration data and train a supervised policy.

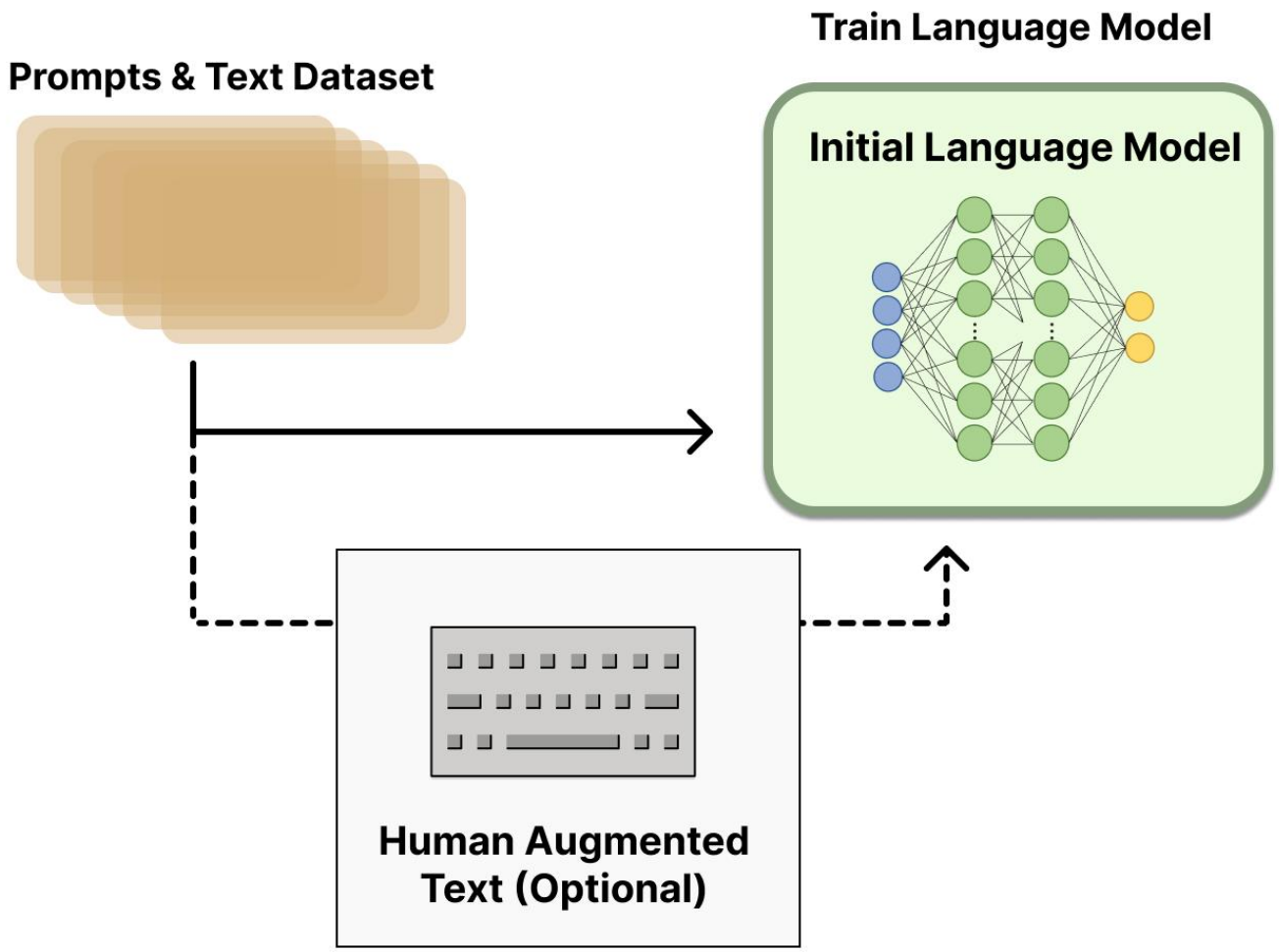
A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.





https://huggingface.co/blog/rlhf

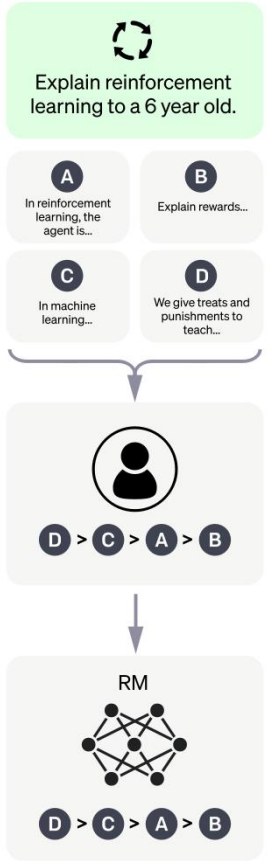
Step2 Reward Modeling

Step 2

Collect comparison data and train a reward model.

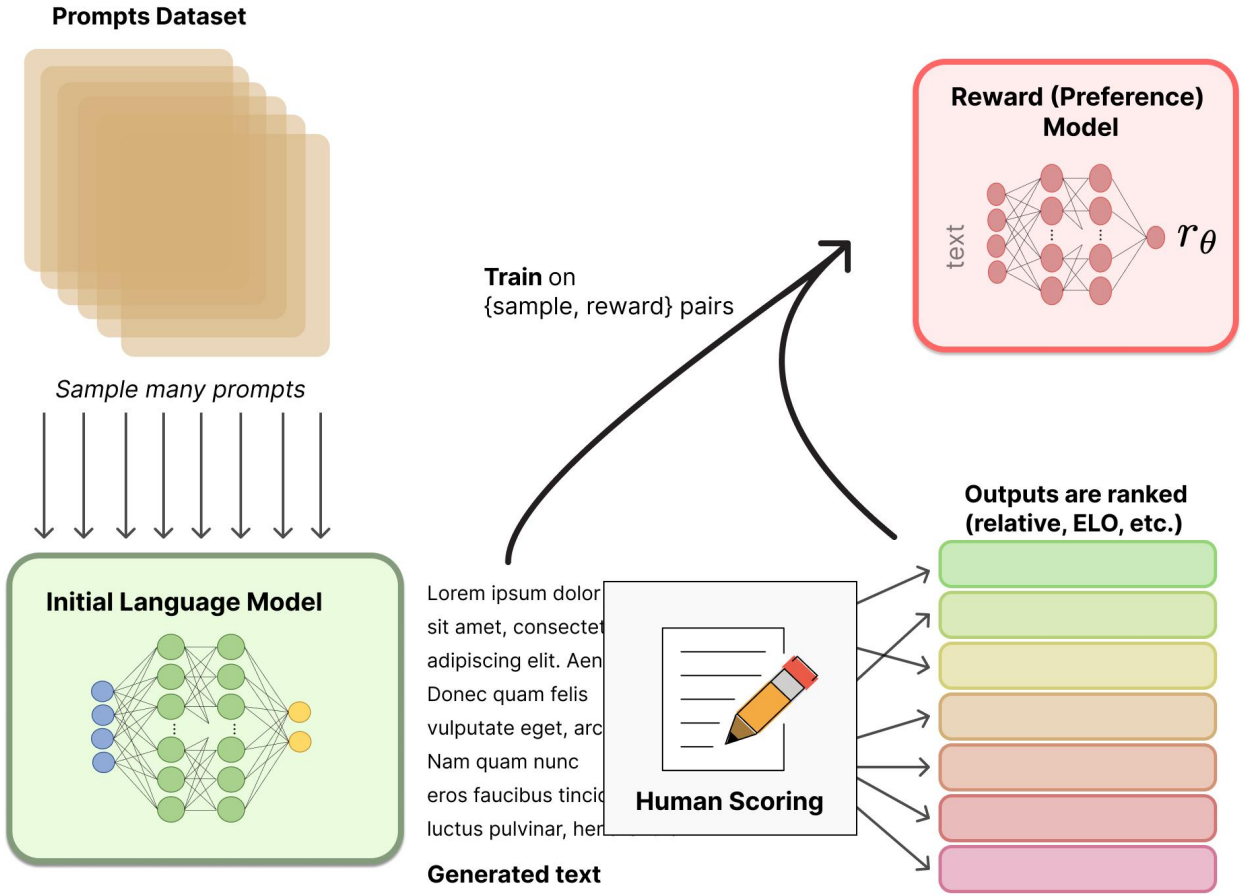
$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta(x,y_w) - r_\theta(x,y_l)\right)\right)\right]$$

A prompt and several model outputs are sampled.



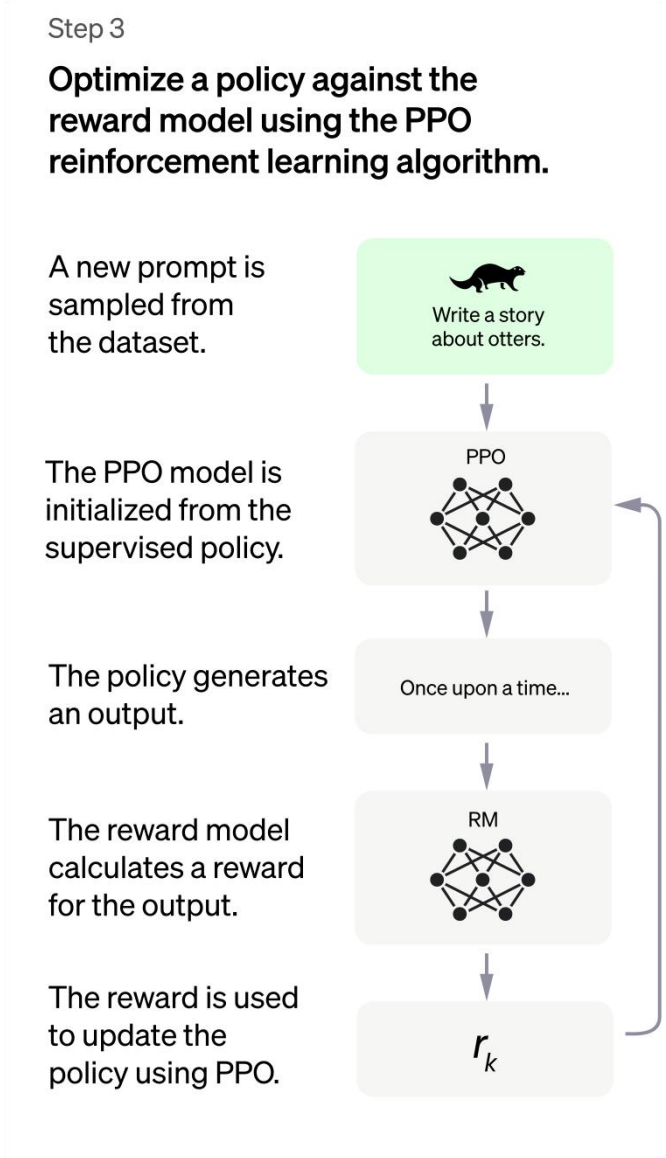
A labeler ranks the outputs from best to worst.

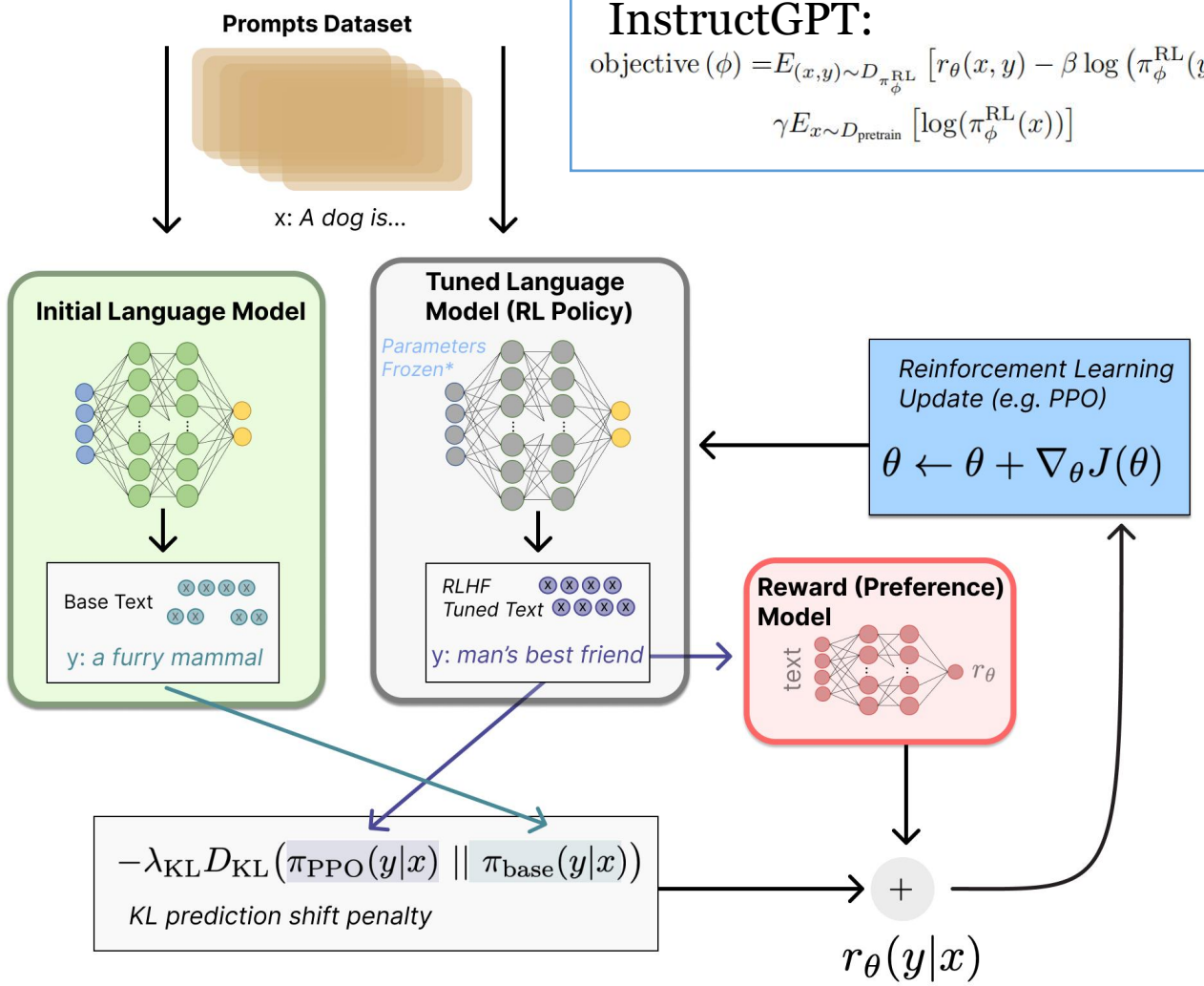
This data is used to train our reward model.



https://huggingface.co/blog/rlhf

Step3 强化学习PPO微调





InstructGPT:

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

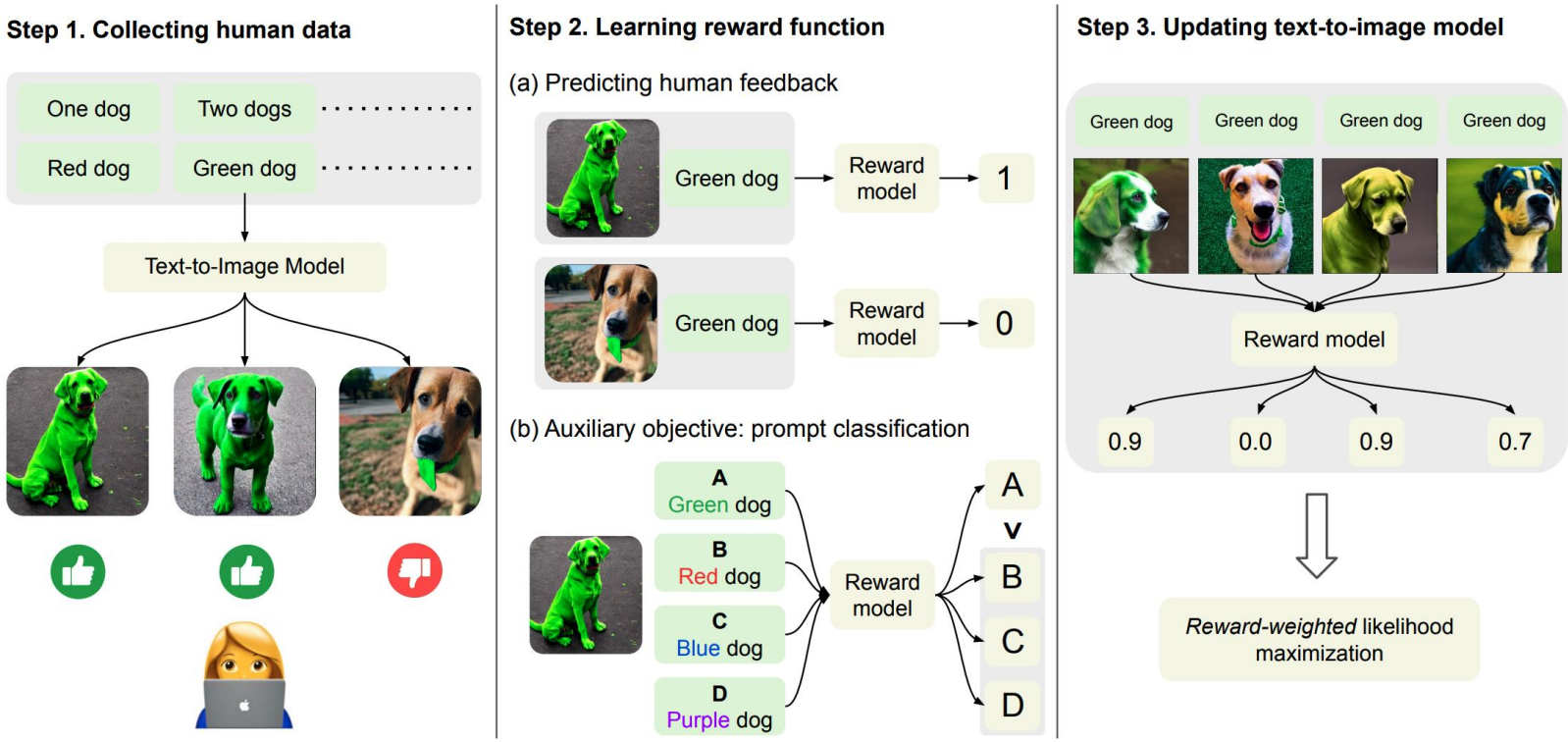
https://huggingface.co/blog/rlhf

目录

狂飙的RLHF：Text-to-Image Diffusion

问题：目前的文生图模型难以生成与输入文本精确匹配的图像，特别是在组合图像生成方面。

贡献：基于人类反馈来精调Stable Diffusion模型来提升生成效果。

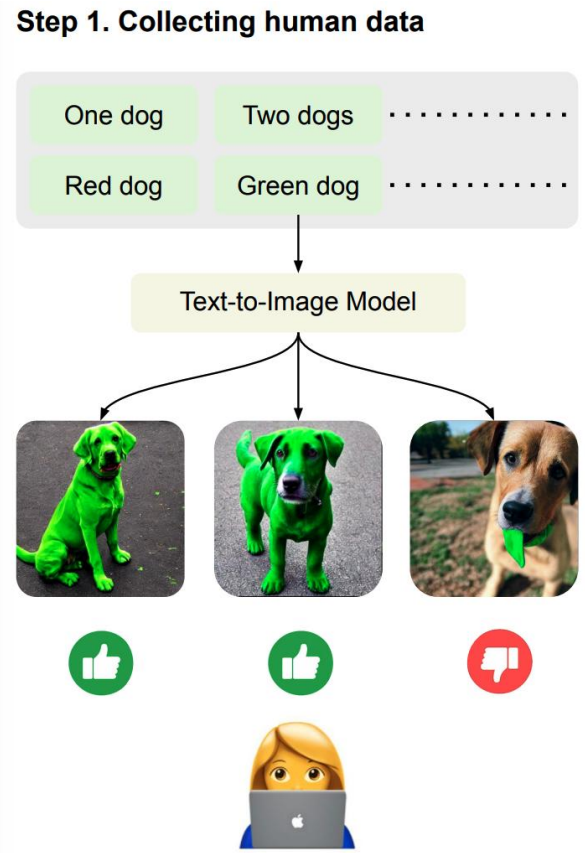


Aligning Text-to-Image Models using Human Feedback, Google, Arxiv 23.02

Step1 收集人类反馈数据

打分二分制：好（1）、差（0）

打分方面：计数、颜色、背景



Category	Examples
Count	One dog; Two dogs; Three dogs; Four dogs; Five dogs;
Color	A green colored dog; A red colored dog;
Background	A dog in the forest; A dog on the moon;
Combination	Two blue dogs in the forest; Five white dogs in the city;

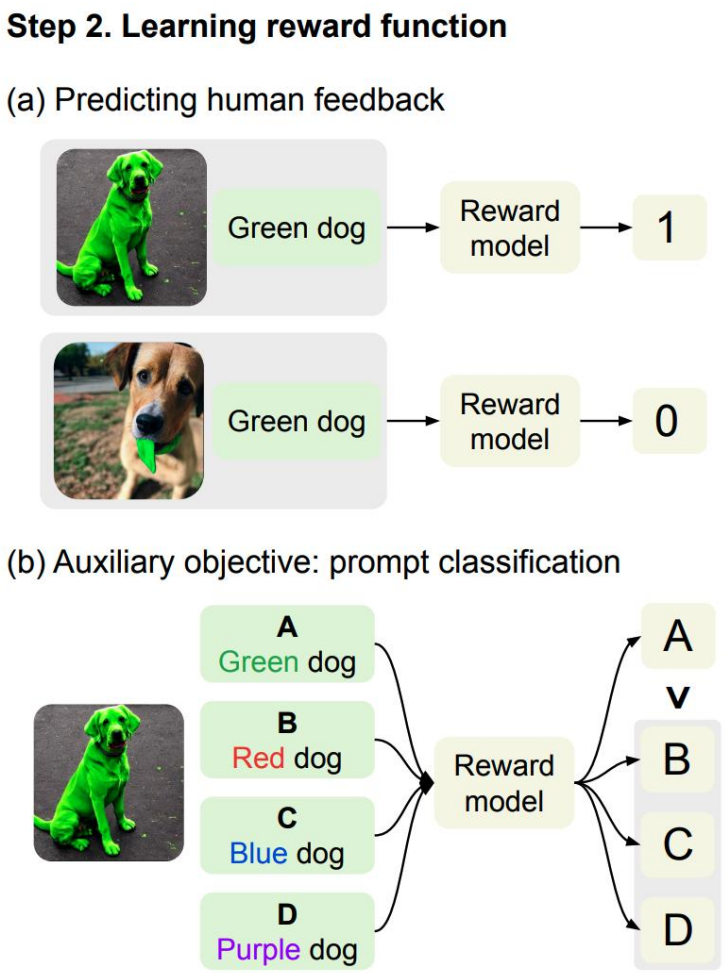
Table 1. Examples of text categories.

Category	Total # of images	Human feedback (%)		
		Good	Bad	Skip
Count	6480	34.4	61.0	4.6
Color	3480	70.4	20.8	8.8
Background	2400	66.9	33.1	0.0
Combination	15168	35.8	59.9	4.3
Total	27528	46.5	48.5	5.0

Table 2. Details of image-text datasets and human feedback.

Aligning Text-to-Image Models using Human Feedback, Googlem, Arxiv 23.02

Step2 学习奖励函数



给定生成的图像和输入的文本，预测生成图像的评分。

1.采用CLIP来提取图像和文本的特征，拼接在一起送入一个2层MLP模型进行评分预测，采用MSE损失来进行训练。

$$\mathcal{L}^{\text{MSE}}(\phi) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{z},y)\sim\mathcal{D}^{\text{human}}} \left[\left(y - r_\phi(\mathbf{x},\mathbf{z})\right)^2 \right].$$

2.辅助任务prompt classification，作为数据增强来提升奖励函数的泛化能力

$$P_\phi(i|\mathbf{x},\{\mathbf{z}_j\}_{j=1}^N) = \frac{\exp(r_\phi(\mathbf{x},\mathbf{z}_i)/T)}{\sum_j \exp(r_\phi(\mathbf{x},\mathbf{z}_j)/T)}, \quad \forall i \in [N],$$

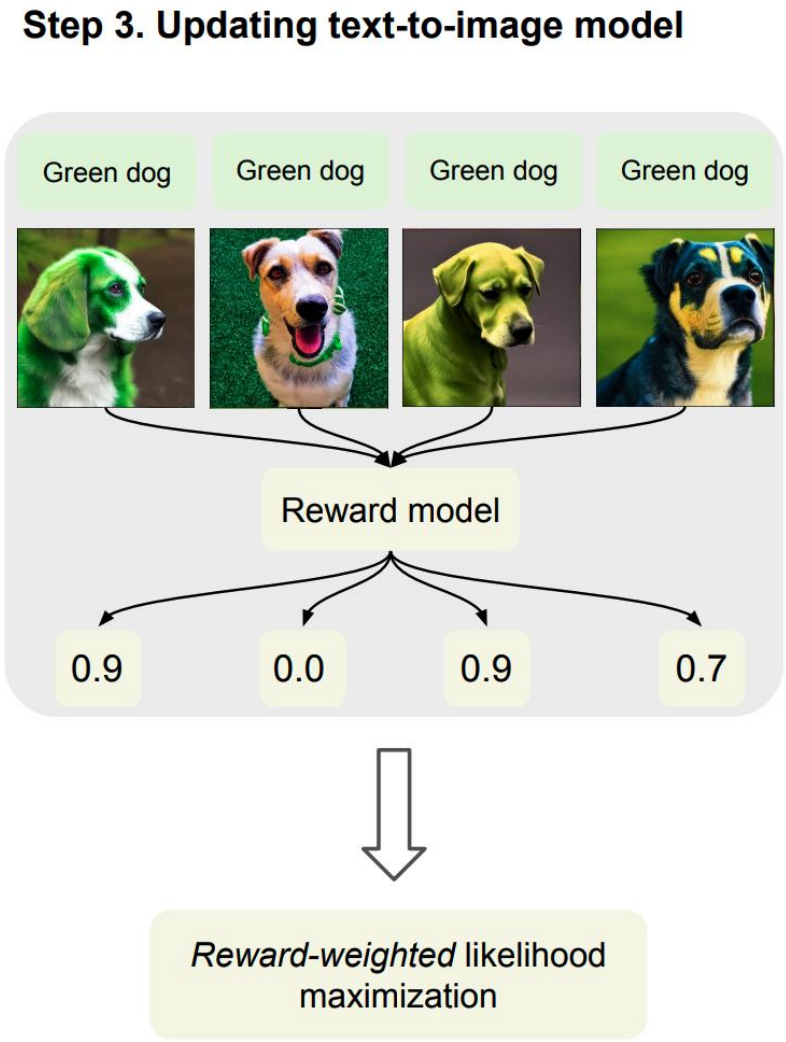
where $T > 0$ is the temperature. Our auxiliary loss is

$$\mathcal{L}^{\text{pc}}(\phi) = \mathop{\mathbb{E}}_{(\mathbf{x},\{\mathbf{z}_j\}_{j=1}^N,i')\sim\mathcal{D}^{\text{txt}}} \left[\mathcal{L}^{\text{CE}}\left(P_\phi(i|\mathbf{x},\{\mathbf{z}_j\}_{j=1}^N), i'\right) \right], (1)$$

$$\mathcal{L}^{\text{reward}}(\phi) = \mathcal{L}^{\text{MSE}}(\phi) + \lambda \mathcal{L}^{\text{pc}}(\phi),$$

Aligning Text-to-Image Models using Human Feedback, Google, Arxiv 23.02

Step3 更新模型



用学习好的奖励函数来精调stable diffusion模型

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{z})\sim\mathcal{D}^{\texttt{model}}} \left[-r_\phi(\mathbf{x},\mathbf{z}) \log p_\theta(\mathbf{x}|\mathbf{z}) \right] + \beta \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{z})\sim\mathcal{D}^{\texttt{pre}}} \left[-\log p_\theta(\mathbf{x}|\mathbf{z}) \right],$$

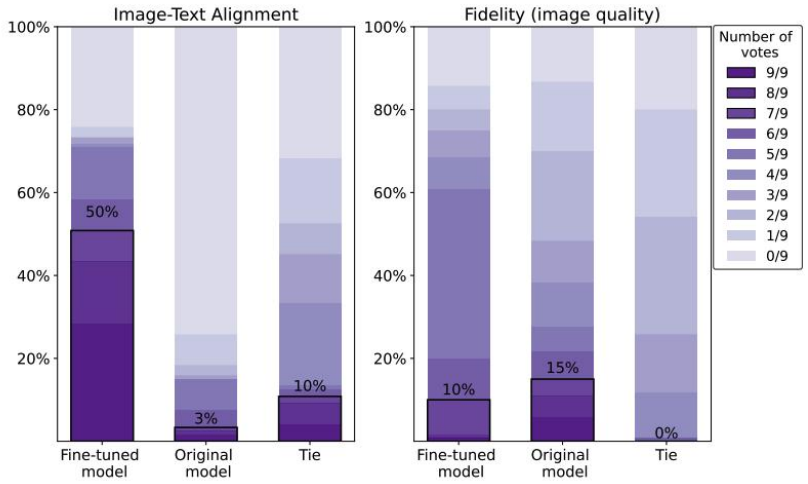
Aligning Text-to-Image Models using Human Feedback, Google, Arxiv 23.02

实验

(a) Seen text prompt: `Two green dogs on the table.`

(b) Unseen text prompt (unseen object): `Four tigers in the field.`

(c) Unseen text prompt (artistic generation): `Oil painting of sunflowers.`



	FID on MS-CoCo (↓)	Average rewards on tested prompts (↑)
Original model	13.97	0.43
Fine-tuned model w.o unlabeled & pre-train	26.59	0.69
Fine-tuned model w.o pre-train	21.02	0.79
Fine-tuned model	16.76	0.79

未来方向

1. More nuanced human feedback：增加人类评分的细粒度，不是简单的0和1打分；
2. Diverse and large human dataset：扩大数据集，提升多样性
3. Different objectives and algorithms：采用RLHF方法

Aligning Text-to-Image Models using Human Feedback,
Google, Arxiv 23.02

总结与展望

现有局限性:

- 仍然可能输出有害或事实上不准确的文本
- 是否可以人工撰写答案?
- 对RLHF的流程、RL 优化器的改进

成功的关键

- 优秀的大模型
- 高质量数据：garbage in garbage out
- Human in the loop:有多少人工就有多少智能

ChatGPT使用技术的发展历程

2017年之前:数学/AI/RL，比如微积分、概率统计、最优化、策略梯度、TRPO算法(PPO的前身，2015年)

2017年6月:RLHF，最后通过TRPO算法迭代策略

2017年7月:PPO算法

2017年6月:Transformer/Self-Attention

2018年6月:GPT

2019年2月:GPT2，融合prompt技术

2020年5月:GPT3，1750B参数

2020年9月:GPT3+RLHF+PPO用于文本摘要

2021年7月:Codex

2021年9月:FLAN大模型,基于指令微调技术Instruction Fine-Tuning (IFT)

2021年第4季度起:OpenAI逐步发展GPT3.5

2022年1月:Google提出思维链技术(Chain of Thought，CoT)

2022年3月:OpenAI正式发布instructGPT，GPT3 + 指令学习 + RLHF + PPO,

2022年11月的ChatGPT：GPT3.5+Codex的代码/推理能力+instruction learning + RLHF + PPO

GPT-4 coming soon...

总结与展望

接口：gpt-3.5-turbo，1000 tokens/$0.002
论文中的使用：

On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective

Jindong Wang[1]*, **Xixu Hu**[1,2‡†], **Wenxin Hou**[3†], **Hao Chen**[4], **Runkai Zheng**[1,5‡], **Yidong Wang**[6], **Linyi Yang**[7], **Wei Ye**[6], **Haojun Huang**[3], **Xiubo Geng**[3], **Binxing Jiao**[3], **Yue Zhang**[7], **Xing Xie**[1]

[1]Microsoft Research, [2]City University of Hong Kong, [3]Microsoft STCA, [4]Carnegie Mellon University, [5]Chinese University of Hong Kong (Shenzhen), [6]Peking University, [7]Westlake University

ChatGPT usage Some authors in this paper are from mainland China where ChatGPT is currently unavailable. In order to conduct this research without disobeying local laws and OpenAI service terms, Hao Chen, who is one of our coauthors and lives in U.S., did all experiments related to ChatGPT and OpenAI. All experiments on ChatGPT are based on its Feb 13 version. Further updates of ChatGPT may lead to change of the results in this paper.

To be continued

02.17 张岚雪 GPTs & ChatGPT

02.24 张琬悦 Instruct Learning

03.03 毕冠群 基于人类反馈的强化学习

03.10 王青悦 ChatGPT的风潮和未来

REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

Thanks

报告人： 毕冠群/时间：2023/3/3