

ChatGPT

狂飙

原理剖析



ZOMI

ChatGPT Talk Overview

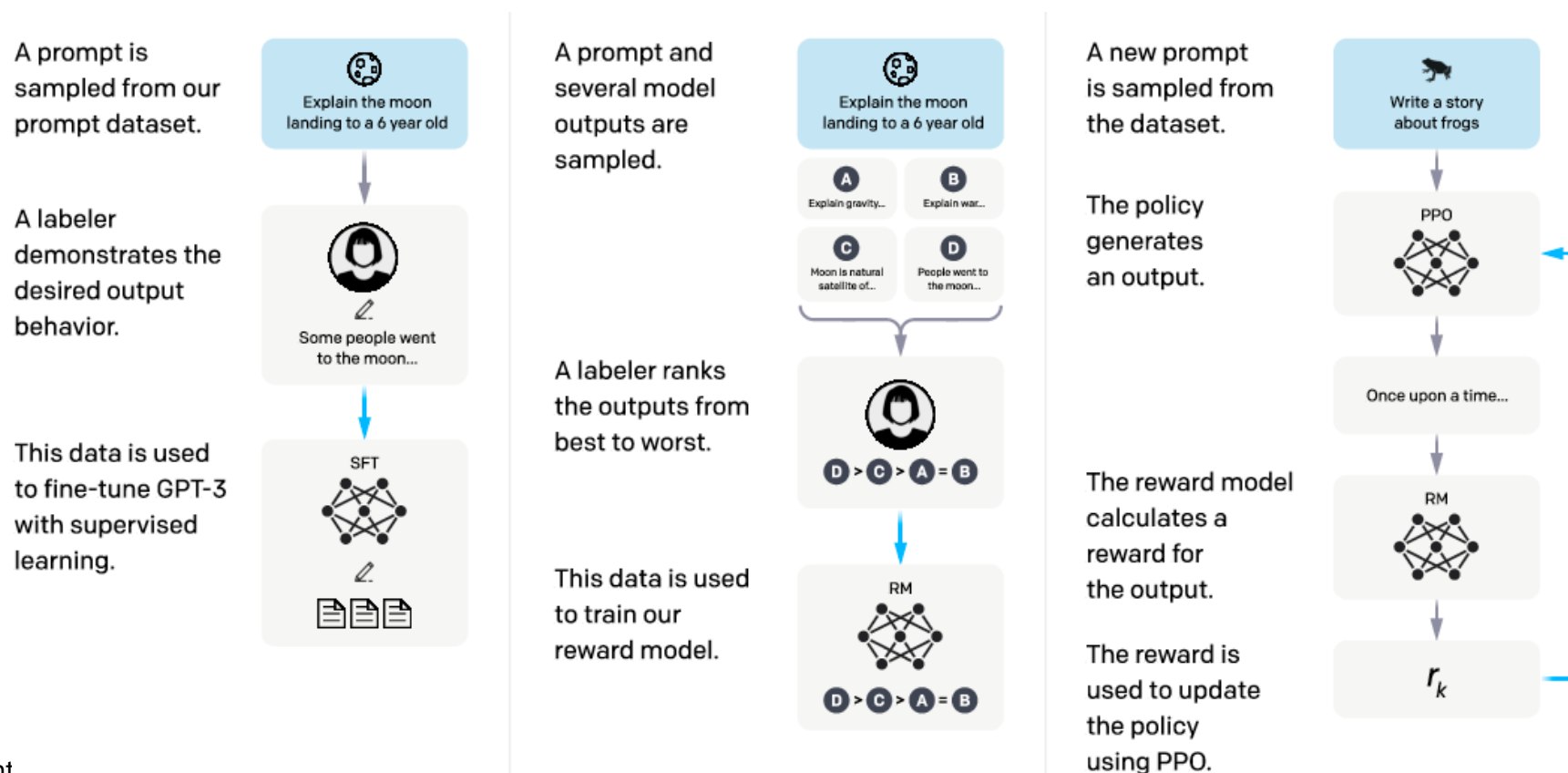
1. BERT 模型与 GPT 模型系列
2. 强化学习加入人类反馈 RLHF 模式
3. 强化学习 PG 和 PPO 算法
4. InstructGPT 原理深度剖析

InstructGPT

原理论文解读

ChatGPT 前身 InstructGPT : 基于 RLHF 微调 GPT-3

- InstructGPT 训练主要分为三个阶段。结合了监督学习和强化学习，先是监督学习让 GPT3 有一个大致的微调方向，然后用强化学习 PPO 算法来更新微调过的GPT3的参数。



ChatGPT 前身 InstructGPT : 基于 RLHF 微调 GPT-3

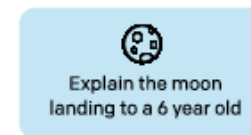
阶段I：利用人类的标注数据（ demonstration data ）去对 GP T3 进行监督训练。

- 1) 先设计了一个prompt dataset，里面有大量提示样本，给出了各种各样的任务描述；
- 2) 其次，标注团队对 prompt dataset 进行标注(本质就是人工回答问题)；
- 3) 用标注后的数据集微调 GPT3（可允许过拟合），微调后模型称为 SFT 模型（ Supervised fine-tuning，SFT），具备了最基本的文本生成能力。

Step 1

Collect demonstration data, and train a supervised policy.

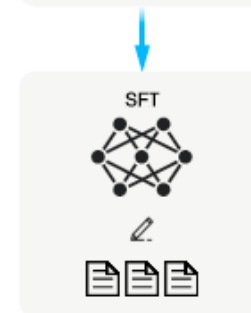
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



ChatGPT 前身 InstructGPT : 基于 RLHF 微调 GPT-3

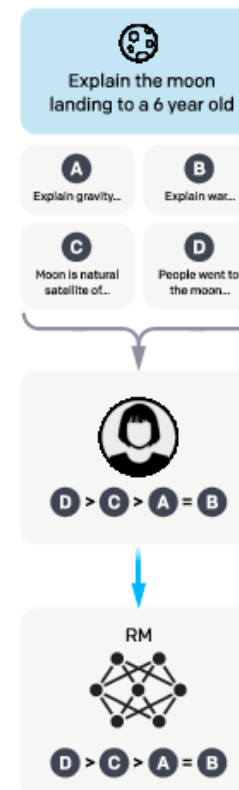
阶段2：通过 RLHF 思路训练奖励模型 RM

- 1) 微调后的 SFT 模型去回答 prompt dataset 问题，通过收集 4 个不同 SFT 输出而获取 4 个回答；
- 2) 接着人工对 SFT 模型生成的 4 个回答的好坏进行标注且排序；
- 3) 排序结果用来训练奖励模型RM (Reward Model)，即学习排序结果从而理解人类的偏好。

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.

ChatGPT 前身 InstructGPT : 基于 RLHF 微调 GPT-3

阶段3：通过训练好的 RM 模型预测结果且通过 PPO 算法优化 SFT 模型的策略。

- 1) 让 SFT 模型去回答 prompt dataset 问题，得到策略的输出，即生成的回答；
- 2) 此时不再让人工评估好坏，而是让阶段 2 RM 模型去给 SFT 模型的预测结果进行打分排序；
- 3) 使用 PPO 算法对 SFT 模型进行反馈更新，更新后的模型称为 PPO-ptx。

Step 3

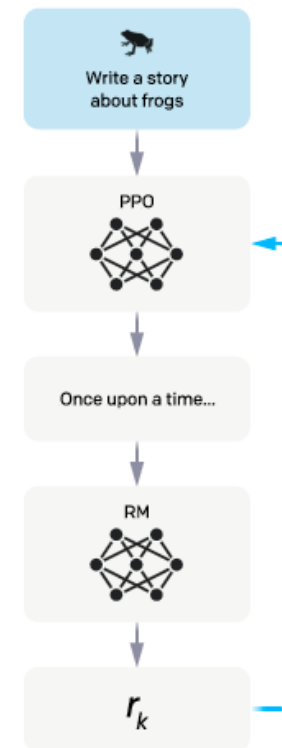
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



InstructGPT

原理深度剖析

阶段2：训练RM模型

- RM 核心是由人类对 SFT 生成的多个输出进行排序，再用来训练 RM。为了让 RM 学到人类偏好（即排序），可以 4 个语句两两组合分别计算 loss 再相加取均值，即分别计算 C_4^2 个损失函数：

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

Step 2

Collect comparison data, and train a reward model.

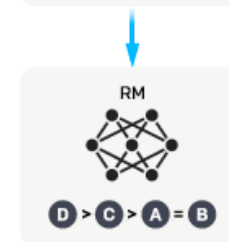
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.

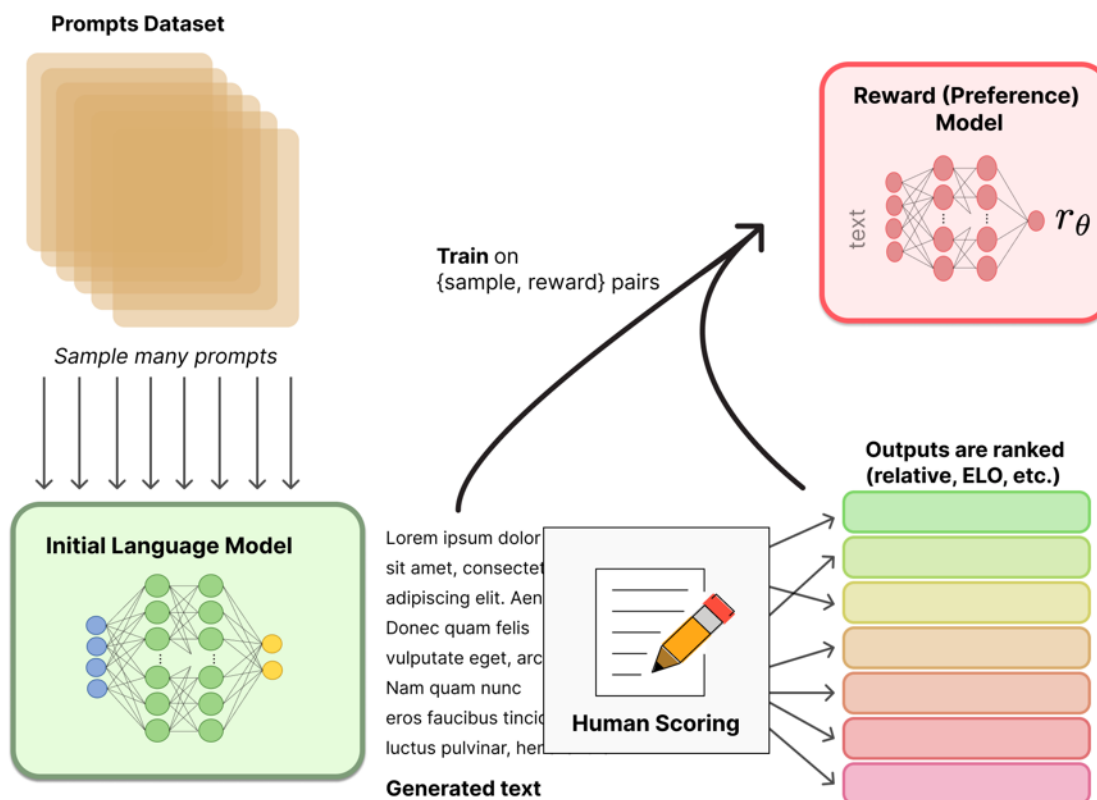


This data is used to train our reward model.



阶段2：训练RM模型

- RM 逐渐学会了给 D 这类语句打高分，给 A/B/C 这类语句打低分，从而模仿人类偏好。所谓 RLFH 人类反馈的强化学习，某种意义上来说，由人类打分来充当 Reward。



阶段3: PPO 优化策略模型

- 强化学习训练的目标函数：

$$\text{objective}(\phi) = E_{(x,y) \sim D_{\pi_{\phi}^{\text{RL}}}} \left[\underbrace{r_{\theta}(x,y)}_{\text{RM 模型}} - \underbrace{\beta \log \left(\pi_{\phi}^{\text{RL}}(y | x) / \pi^{\text{SFT}}(y | x) \right)}_{\text{KL 散度对比策略模型 RL 和 SFT}} \right] + \underbrace{\gamma E_{x \sim D_{\text{pretrain}}} \left[\log(\pi_{\phi}^{\text{RL}}(x)) \right]}_{\text{修正偏置项}}$$

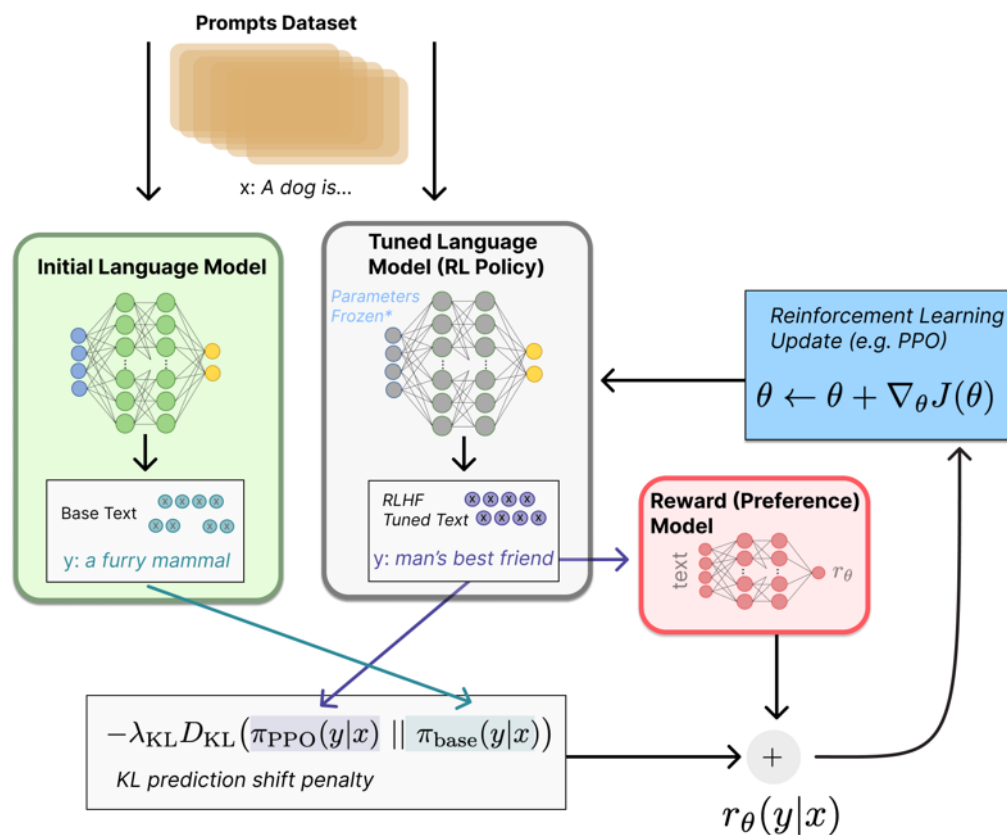
RM 模型

KL 散度对比策略模型 RL 和 SFT

修正偏置项

阶段3: PPO 优化策略模型

- 迭代式的更新奖励模型 RM 和策略模型 SFT，让奖励模型 RM 对 PPO 模型输出质量的刻画愈加精确，并使得输出文本变得越来越符合人的认知。



思考

思考

- 大模型通过计算的方式模拟人类的思考，类似于ChatGPT的RLHF技术是否会给世界带来新的技术产业革命？
- ChatGPT 使用了 Ray 作为细粒度的并行计算和异构计算，管理分配 RL/DL 模型完成复杂训练任务，方便利用强化学习对环境和计算任务进行控制，这对AI框架的分布式能力边界带来哪些新的冲击？
- ChatGPT 非常重数据交互，存算一体技术会不会针对ChatGPT等应用出现专用芯片和新的架构？



引用

1. Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018)
2. Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9
3. Dale, Robert. "GPT-3: What's it good for?." Natural Language Engineering 27.1 (2021): 113-118
4. Floridi, Luciano, and Massimo Chiriatti. "GPT-3: Its nature, scope, limits, and consequences." Minds and Machines 30 (2020): 681-694.
5. Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901
6. Ouyang, Long, et al. "Training language models to follow instructions with human feedback." arXiv preprint arXiv:2203.02155 (2022)
7. <https://openai.com/blog/instruction-following/>
8. <https://openai.com/blog/chatgpt/>
9. <https://sh-tsang.medium.com/review-instructgpt-training-language-models-to-follow-instructions-with-human-feedback-7fce4bf9059a>
10. <https://jalammar.github.io/how-gpt3-works-visualizations-animations/>
11. <https://jalammar.github.io/>



BUILDING A BETTER CONNECTED WORLD

THANK YOU

Copyright©2014 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.