# 大模型（LLMs）强化学习—— PPO 面

来自： AiGC面试宝典

宁静致远

2024年01月27日 20:47





## 一、大语言模型RLHF中的PPO主要分哪些步骤?

大语言模型RLHF中的PPO 分为：

1. 采样
2. 反馈
3. 学习

对应的实现逻辑如下：

```
policy_model = load_model()

for k in range(20000):
    # 采样（生成答案）
    prompts = sample_prompt()
    data = respond(policy_model, prompts)

    # 反馈（计算奖励）
    rewards = reward_func(reward_model, data)

    # 学习（更新参数）
    for epoch in range(4):
        policy_model = train(policy_model, prompts, data, rewards)
```

## 二、举例描述一下 大语言模型的RLHF?

**大语言模型的RLHF，实际上是模型先试错再学习的过程。**

大语言模型的RLHF 好比是：老师与学生的角色

- 我们扮演着老师的角色，给出有趣的问题。模型则会像小学生一样，不断尝试给出答案。
- 模型会根据我们给出的问题，写出它觉得正确的答案，但是这些答案不一定是真的答案，需要我们结合正确答案进行打分。如果它表现得好，就会给予它高声赞扬；如果它表现不佳，我们则会给予它耐心的指导和反馈，帮助它不断改进，直到达到令人满意的水平。

## 三、大语言模型RLHF 采样篇

### 3.1 什么是 PPO 中 采样过程?

PPO 中 采样过程：学生回答问题的过程，是模型根据提示（prompt）输出回答（response）的过程，或者说是模型自行生产训练数据的过程。

> eg:

| prompt | response |
| --- | --- |
| 请告诉我三种常见的动物。 | 猫，狗，鹦鹉。 |
| 如何评价电影《爱乐之城》？ | 音乐的经典令人赞叹不已，结局却让人感到五味杂陈。 |
| 詹姆斯和库里谁更伟大? | 他们都很伟大，我无法比较。 |

### 3.2 介绍一下 PPO 中 采样策略?

PPO 中 采样工作 通过一种**策略（policy）：policy由两个模型组成，一个叫做演员模型（Actor），另一个叫做评论家模型（Critic）。它们就像是学生大脑中的两种意识，一个负责决策，一个负责总结得失。**

> 演员：我们想要训练出来的大模型。在用PPO训练它之前，它就是RLHF的第一步训练出来的SFT（Supervised Fine-Tuning）model。输入一段上下文，它将输出下一个token的概率分布。
>
> 评论家：强化学习的辅助模型，输入一段上下文，它将输出下一个token的“收益”。

### 3.3 PPO 中 采样策略中，如何评估“收益”?

从下一个token开始，模型能够获得的总奖励（浮点数标量）。这里说的奖励包括Reward Model给出的奖励。