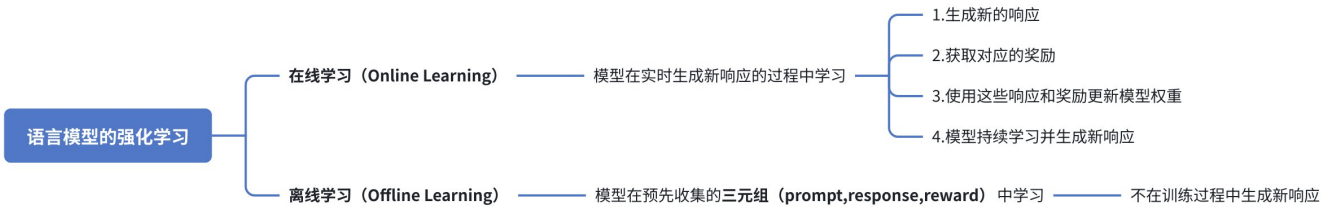
Online Reinforcement Learning

1. 语言模型的强化学习:



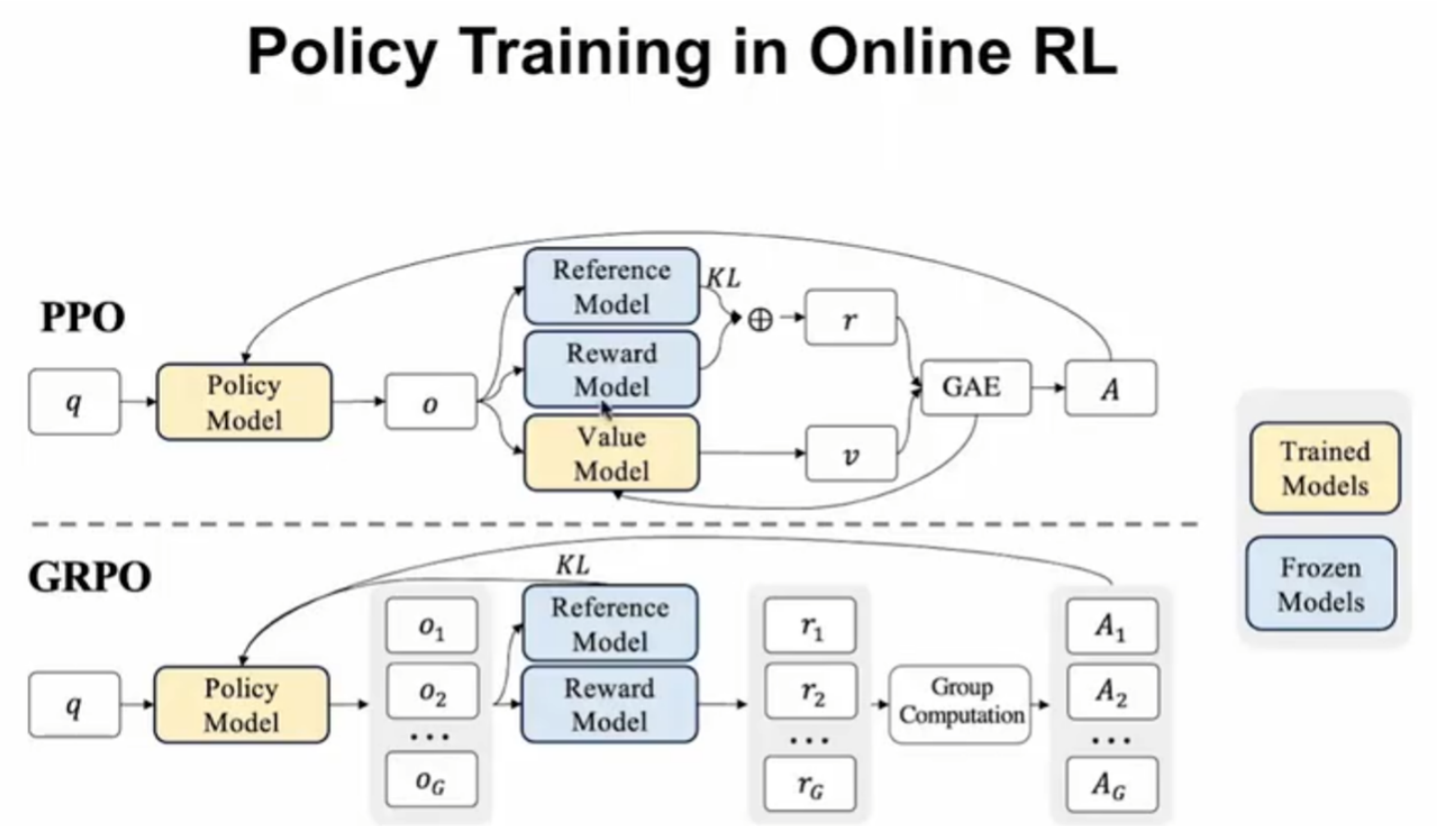
2. 在线强化学习工作机制：让模型自主探索更好的响应，流程如下
   a. 准备一批Prompt
   b. 将这批Prompt输入模型
   c. 模型生成相应的Response
   d. 将（prompt,response）对输入**奖励函数（Reward Function）**
   e. 奖励函数为每对（prompt,response）打分
   f. 获得(prompt,response,reward）三元组
   g. 使用这些数据来更新语言模型

   更新语言模型的方法很多，本课重点为**PPO（Proximal Policy Optimization 近端策略优化）**和**GRPO（Group Relative Policy Optimization 群体相对策略优化）**
3. 奖励函数（Reward Function）：

训练好的奖励模型 （Reward Model）	• 收集多个模型响应，由人类标注选择更优的响应 • 使用这些人类偏好数据训练奖励模型 • 奖励模型通过优化以下损失函数学习： $$L = \log(\sigma(r_j - r_k))$$ ◦ 若人类认为响应j优于k，则鼓励模型提升r_j，降低r_k	特点： • 通常基于已有的Instruct模型优化 • 通过大规模人类或机器生成偏好数据训练 • 可应用于开放式任务，如聊天能力、安全性提升等 • 在“正确性导向”的任务，如代码、数学、函数调用中可能不够精确
	更适用于“正确性导向”任务	特点：

可验证性奖励（Verifiable Reward）	• 数学任务：验证输出是否与标准答案匹配 • 编程任务：通过**单元测试（Unit Tests）**检验代码执行结果是否正确	• 需提前准备**真值（Ground Truth）**或测试集 • 准备成本较高，但奖励信号更精确可靠 • 更适合训练推理类模型（Reasoning Models），如代码、数学领域

4. PPO VS GRPO



方法	工作流程	总结
PPO（Proximal Policy Optimization）	1. 输入一组**查询（queries）（q）** 2. 通过**策略模型（Policy Model）**（语言模型本身）生成响应 3. 相应被送入以下3个模块： a. **参考模型（Reference Model）**：计算KL散度，限制模型不偏离原始分布 b. **奖励模型（Reward Model）**：计算奖励 c. **价值模型（Value Model）**或**评价者模型（Critic Model）**：为每个token分配价值	PPO的目标函数： $J_{PPO}(\theta)=E_{q\sim P(Q),o\sim\pi_{\theta_{old}}(O\mid q)}\left[\frac{1}{\lvert o\rvert}\sum_{t=1}^{\lvert o\rvert}\min\left[\frac{\pi_\theta(o_t\mid q,o_{<t})}{\pi_{\theta_{old}}(o_t\mid q,o_{<t})}A_t,\text{clip}\left(\frac{\pi_\theta(o_t\mid q,o_{<t})}{\pi_{\theta_{old}}(o_t\mid q,o_{<t})},1-\varepsilon,1+\varepsilon\right)A_t\right]\right]$ • 每个token拥有独立的优势值 • 反馈粒度更细 • 需额外训练价值模型——>占用更多GPU内存

	4. 使用**广义优势估计（Generalized Advantage Estimation,GAE）**来计算每个token的**优势函数（Advantage）**，反映该token的贡献	
GRPO（Group Relative Policy Optimization）	1. 对每个prompt模型生成多个响应（*O1，O2，...，Og*） 2. 对每个响应计算**奖励（Reward）和参考模型的KL散度** 3. 对同一组（Group）响应计算**相对奖励（Relative Reward）** 4. 将相对响应作为整个响应的优势值 5. 使用此优势更新策略模型	• 不再需要**价值模型（Value Model）** • 所有token在同一响应中共享相同优势值 • 更节省显存，但优势估计较粗糙

特征	PPO	GRPO
优势估计	基于价值模型 (Value Model) 的精细估计	基于响应组的相对奖励 (Relative Rev
计算粒度	每个 Token 拥有独立优势	整个响应共享同一优势
显存需求	较高（需训练 Critic）	较低（无 Critic）
样本效率	高（样本利用率好）	较低（需更多样本）
奖励适配	适合连续或模型化奖励	适合二元/可验证奖励
应用场景	聊天、对齐、安全优化	数学、代码、推理任务

5. 代码解析

在线强化学习（Online RL）实践.md