1.引言

自 2022 年 OpenAI 发布 ChatGPT 以来，大语言模型（Large Language Model, LLM）迅速崛起为人工智能领域的核心研究方向，引发了广泛关注与深度应用。其中，强化学习（Reinforce Learning, RL）扮演着关键角色 —— 其最初通过人类偏好数据的运用，有效确保了大语言模型生成内容的道德合规性，并使其与人类价值观保持一致。而 2025 年初，深度求索（DeepSeek）发布的开源推理模型 Deepseek-R1，更是将强化学习的受关注度推向了新的高度。

2.基本 RL 算法

从是否采用参数化策略的角度划分，强化学习算法可清晰地归为两大阵营：一类是以 Q-learning 为典型代表的 value-base 算法，这类算法核心在于通过学习动作价值函数（Value Function）来指导决策，即通过评估不同动作的预期收益确定最优策略；另一类则是以 Reinforce 为标志性算法的 policy-base 算法，其直接对策略本身进行参数化建模，通过优化策略参数来最大化累积奖励。

在大语言模型（LLM）的训练场景中，由于需要处理高维度的离散动作空间（如文本序列的生成），且对策略的表达灵活性要求更高，**policy-base 算法凭借其直接优化策略的特性更能适配此类场景**，因此成为主流选择。本文将聚焦于 policy-base 算法体系展开深入探讨，以下将从经典的 Reinforce 算法出发，按发展脉络依次介绍一系列具有里程碑意义的改进算法。

2.1. Reinforce 算法

Reinforce 算法 作为 policy-base 算法的奠基之作，其核心思想是通过蒙特卡洛采样估计策略梯度，直接更新策略参数。

$$J(\theta) \approx E_{\pi_\theta} \sum_{t=0}^{T-1} \sum_{a} \pi_\theta(a|s_{i,t}) G_{i,t}$$

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) G_{i,t}$$

其中，$J(\theta)$为策略目标函数（累积期望奖励），θ为策略参数，$\pi_\theta(a|s)$为状态s下采取动作a的概率，$G_{i,t}$为第i条轨迹中时刻t的累积回报（从t到终止状态的奖励总和），N为采样轨迹数量。 作为蒙塔卡罗方法，Reinforce可能具有较大的方差，收敛较慢。

2.2. TRPO（Trust Region Policy Optimization）

TRPO为解决 Reinforce 策略更新波动过大的问题，TRPO 引入“信任域”约束，确保策略更新在安全范围内。（本质上是延续了保守策略迭代的思想，在理论上证明了策略π_θ逐步改进）

$$\max_\theta \mathbb{E}_{s\sim\rho_{\pi_{\text{old}}}} \left[\frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)} A_{\pi_{\text{old}}}(s,a) \right]$$
$$\text{s.t. } \mathbb{E}_{s\sim\rho_{\pi_{\text{old}}}} \left[D_{\text{KL}}(\pi_{\text{old}}(\cdot|s) \| \pi_\theta(\cdot|s)) \right] \le \delta$$

其中，π_{old}为旧策略，$A_{\pi_{\text{old}}}(s,a) = Q_{\pi_{old}}(a,s) - V_{\pi_{old}}(s)$为优势函数（衡量动作$a$相对平均收益的优势），$D_{\text{KL}}$为 KL 散度（衡量新旧策略的差异），$\delta$为信任域半径。

理论补充：从理论层面看，TRPO 的理论依据其实也支持用**惩罚项（penalty）替代硬约束**，即将问题转化为无约束优化：

$$\underset{\theta}{\text{maximize}} \quad \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \hat{A}_t - \beta \mathrm{KL}\left[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)\right] \right]$$

其中 β 是惩罚系数。该思路源于：某种替代目标（通过“状态上的最大 KL 散度”而非“均值”计算），能为策略 π_θ的性能提供**悲观下界（pessimistic bound）**。但实际中，TRPO 更倾向用硬约束而非惩罚项，原因在于**惩罚系数 β 难以选择——**不同问题（甚至同一问题学习过程中环境特性变化时），很难找到一个普适的 β 保证效果。　若想设计一阶算法（如用 SGD）模拟 TRPO 的单调性能提升，仅简单固定 β 优化是不够的，还需额外修改（这也为后续 PPO 等算法的诞生埋下伏笔 ）。[参考PPO论文的叙述]

2.3. PPO（Proximal Policy Optimization）

PPO通过简化 TRPO 的约束条件，以更高效的方式实现稳定更新，成为目前应用最广泛的 policy-base 算法。为与LLM更为贴近，我们介绍在RLHF中最为常见的Actor-Critic Style PPO 。

$$J_{\text{CLIP}}(\theta) = \mathbb{E}_{s,a\sim\pi_{\text{old}}} \left[\min \left(\frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)} A_{\pi_{\text{old}}(s,a)}(s,a), \text{clip} \left(\frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)}, 1-\epsilon, 1+\epsilon \right) A_{\pi_{\text{old}}(s,a)} \right) \right]$$

其中，ϵ为截断系数（通常取 0.1 或 0.2），clip函数将策略比值限制在$[1-\epsilon, 1+\epsilon]$范围内，以防止造成较大的策略梯度，保证策略迭代的稳定性。

总的来说，PPO with Actor - Critic 实现了两项关键改进，即 clip 机制与 Actor - Critic 框架。其中，clip 机制尤为核心，现已成为后续算法的普遍共识；而 Actor - Critic 框架在大语言模型（LLM）的后续应用里，在Reinforce++、RLOO、GRPO等算法中却逐渐被弃用。主要原因在于，LLM 应用场景中，反馈多依赖**结果奖励函数**，依靠这类函数难以训练出 token 级别的 Critic 模型 。

2.4. GRPO（Group Relative Policy Optimization，组相对策略优化）

GRPO算法首次在DeepSeek-Math文章中提出，在2025年初DeepSeek-R1开源，这一算法引起广泛关注。GRPO 针对大语言模型（LLM）强化学习场景中价值函数训练难题，核心是用**组相对奖励与采样平均基线**，规避 PPO 对 Critic 模型的依赖。

对每个问题 q，GRPO 从旧策略 $\pi_{\theta_{\text{old}}}$采样一组输出 $\{o_1, o_2, \cdots, o_G\}$，通过最大化**组相对优化目标**更新策略 π_θ，目标函数为：

$$J_{\text{GRPO}}(\theta) = \mathbb{E}\left[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)\right]$$
$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t},\ \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{I}$$

其中 ϵ, β：超参数ϵ 控制策略更新截断范围，β正则 KL 散度 ；　$\hat{A}_{i,t}$：**组内相对优势**，仅基于同组输出的相对奖励计算，体现当前 token 在组内的“相对价值”；　$\mathbb{D}_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right]$：策略 π_θ与参考策略的 KL 散度，直接加入损失作正则，避免干扰优势计算。　在 GRPO 框架中，使用以下无偏估计，计算正则策略更新的 KL 散度（Kullback - Leibler Divergence），保证KL为正 。

$$\mathbb{D}_{KL}\left[\pi_\theta \| \pi_{\text{ref}}\right] = \frac{\pi_{\text{ref}}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})} - \log\frac{\pi_{\text{ref}}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})} - 1$$

	PPO（ACTOR - CRITIC）	GRPO
价值函数依赖	需训练逐 token 价值函数（Critic）	无独立价值函数，用**组采样平均奖励**作基线
优势计算逻辑	基于绝对价值函数与奖励的差值（(A = r - V(s))）	基于**组内输出相对奖励**（仅同组样本对比）
LLM 适配性	难处理“仅最后 token 奖励”场景（价值函数训练复杂）	天然适配结果奖励，无需逐 token 价值拟合

2.5. DAPO（Decoupled Clip and Dynamic Sampling Policy Optimization）

DAPO 是字节提出的方法，创新性地将“截断机制”与“动态采样”解耦，进一步提升策略适应复杂数据分布的能力。

论文中提到主要有四点改进：

1. **Clip - Higher**：解耦PPO中clip参数，鼓励模型探索，促进系统多样性，避免**熵坍缩**；（目前基于熵的研究可以说非常丰富，与DAPO观点略有不同，普遍认为LLM在后训练过程中不具备再探索能力，RL过程中策略熵逐渐降低，需要的方法控制策略熵，从而最大化奖励。通俗理解就是RL的过程是一个熵换取奖励的过程，良好的熵控制能够最大化奖励。下篇文章准备讨论下）
2. **Dynamic Sampling**：针对GRPO组相对奖励，如果获取的reward相同，对造成梯度消失，因此拒绝该样本，从而提升训练效率与稳定性
3. **Token-Level Policy Gradient Loss**：针对在长思维链（long-CoT ）强化学习场景中，sample-level的梯度计算导致长思维链token权重较低，因此提出该方法。
4. **Overlong Reward Shaping**：奖励模型倾向于较长的回答，减少奖励噪声，对过长的response做约束。

策略优化目标函数

$$J_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}$$
$$\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta),1-\varepsilon_{\text{low}},1+\varepsilon_{\text{high}}\right)\hat{A}_{i,t}\right)\right]$$

$$\text{s.t. } 0 < \left|\{o_i \mid \text{is\_equivalent}(a, o_i)\}\right| < G,$$

其中$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$，　$\hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^{G})}{\text{std}(\{R_i\}_{i=1}^{G})}$.

注：在DAPO算法中，移除了 RLHF 中目标策π_θ与参考模型π_{ref} 的KL正则化项，作者认为在训练长CoT推理模型时，模型分布可能与初始模型有明显差异，因此没有必要进行这种限制。（后续有文章认同该改进）

3.总结

以上是笔者对LLM中主流在线强化学习的简单梳理，借助了一些AI工具并融入了自己的理解，如有理解不到位的地方恳请指正。第一篇文章多多见谅！