第 7 章

大模型对齐中的 PPO、GRPO 与 DPO

一、引言 · 二、PPO：Online RLHF 的标准基准

一、引言

1.1 LLM 对齐：为什么需要 RL？

大型语言模型（LLM）经过大规模预训练后，具备了强大的语言理解和生成能力。但预训练的目标是**预测下一个 Token**，这并不等同于"生成对人类有用、安全、诚实的回答"。模型可能会编造事实、生成有害内容、或者答非所问。

> **对齐（Alignment）**的目标就是弥合这一鸿沟——让模型的行为符合人类的意图和价值观。

标准的对齐流程通常分为两步：

1 SFT

用高质量的人工标注数据对模型进行有监督微调，教会模型"对话的格式"。

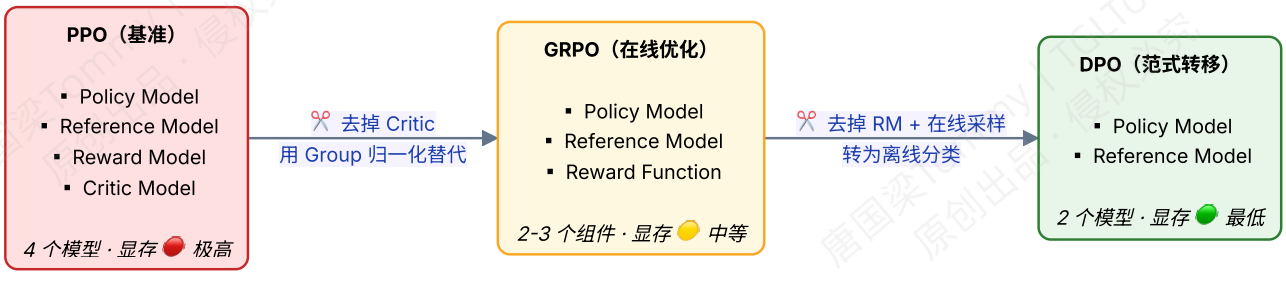
2 RLHF

用 RL 算法进一步优化模型，使其生成更符合人类偏好的回答。

本章聚焦第二步，深入讲解三个核心算法：**PPO、GRPO、DPO**。

1.2 本章主线："架构做减法"

本章不是三个算法的简单罗列，而是一个**"从复杂到简单"的技术演进故事**。核心线索是：**每一个新算法都在尝试减少训练所需的模型组件，降低工程复杂度和资源消耗**。



三大算法的"减法"演进路线

PPO 标准 Online RLHF 的基准方案，功能强大但"重"

GRPO Critic 真的必要吗？用 Group 采样替代，专攻推理任务

DPO Reward Model 和在线采样真的必要吗？直接从偏好数据学习

二、PPO：Online RLHF 的标准基准

2.1 核心思想

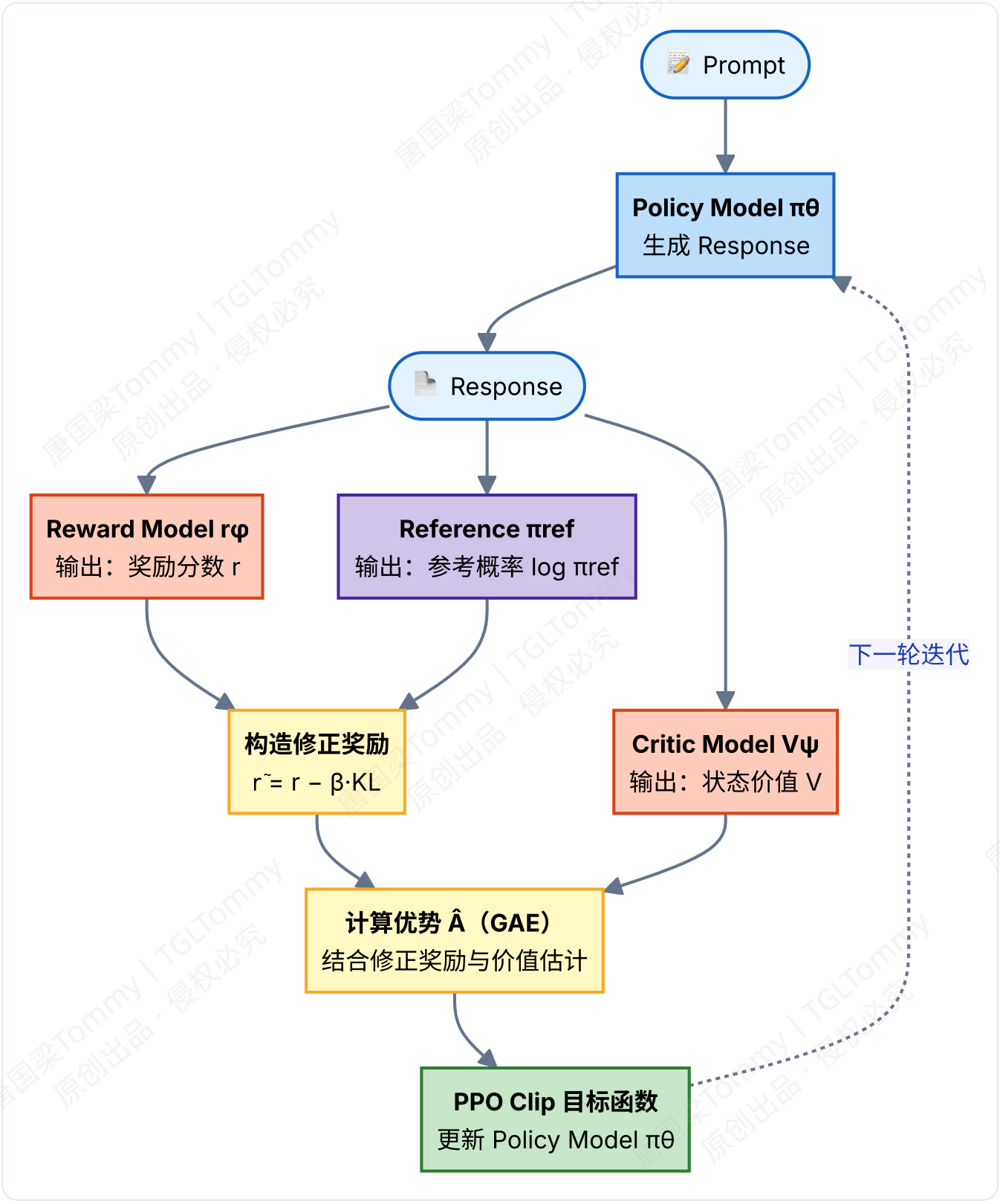
PPO 是基于 Actor-Critic 架构的策略优化算法，通过 clip 机制限制更新幅度来保证训练稳定性。在 LLM 对齐中，它是 RLHF 的经典实现方案——功能最完整，但也最"重"。

PPO 在 LLM 对齐中需要 **4 个模型**协同工作，可以类比为一个"驾校教练系统"：

模型	符号	角色类比	职责
Policy Model	π_θ	学员	生成回答，是唯一被优化的目标模型
Reward Model	r_φ	考官	对回答打分，提供人类偏好信号
Critic Model	V_ψ	副教练	预测学员表现水平，帮助计算"比预期好多少"（优势）
Reference Model	π_{ref}	规则手册	约束学员不要跑偏太远（KL 惩罚）

唐国梁Tommy | TGLTommy 原创出品 · 侵权必究

2.2 架构与数据流



PPO 在 LLM 对齐中的完整训练流程

1. **采样阶段**：Policy Model 针对一批 Prompt 生成 Response（在线采样）
2. **打分阶段**：Reward Model 对每个 (Prompt, Response) 对打分
3. **奖励修正**：将 RM 打分减去 KL 惩罚，得到修正奖励，防止策略偏离 Reference 太远

4 **优势估计**：Critic Model 预测状态价值，结合修正奖励，用 GAE 计算每个 Token 的优势

5 **策略更新**：用 PPO Clip 目标函数更新 Policy Model 参数

2.3 关键公式

PPO Clip 目标函数：

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t \right) \right]$$

其中 $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ 是新旧策略的概率比。

$r_t > 1$

新策略更倾向于生成这个 Token

$r_t < 1$

新策略在回避这个 Token

裁剪机制将 r_t 限制在 $[1-\varepsilon, 1+\varepsilon]$ 内，当策略变动太大时，抑制目标函数的增长，防止策略更新过猛。ε 是裁剪参数（如 $\varepsilon = 0.2$）。

优势函数 $\hat{A}_t$

告诉策略当前选择的动作 a，比该状态下的平均水平好多少：

- `A > 0` 这个动作好，鼓励策略增加它的概率
- `A < 0` 这个动作不好，惩罚它，降低概率
- `A = 0` 这个动作刚好平均，无需改变

2.4 PPO 联合损失函数

PPO 同时优化三件事：① 提升策略质量；② 学准价值预测；③ 保持探索多样。

$$L^{CLIP+VF+S}(\theta, \psi) = \hat{\mathbb{E}}_t \left[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\psi) + c_2 S[\pi_\theta](s_t) \right]$$

损失项	公式	作用	直觉
策略损失 L^{CLIP}	Clip 目标函数	更新 Policy 参数，鼓励优势更大的动作	"写更好的答案"
价值损失 L^{VF}	$(V_\psi(s_t) - R_t)^2$	更新 Critic 参数，让价值预测逼近真实回报	"打分更准确"
熵奖励 $S[\pi]$	$-\sum_a \pi_\theta(a\|s_t) \log \pi_\theta(a\|s_t)$	保持策略分布足够随机，避免过早收敛	"多试试别的写法"

唐国梁Tommy | TGLTommy 原创出品 · 侵权必究

c_1 控制价值损失权重（通常取 0.5），c_2 控制熵奖励权重（通常取 0.01）。在 LLM RLHF 中，Policy 和 Critic 通常是**独立的模型**（不共享参数）。

2.5 LLM 场景下的特殊性

维度	传统 RL	LLM-RLHF
动作空间	离散有限（如上下左右）或连续低维	超大离散空间（词表大小，通常 32K-150K）
序列长度	通常较短（几十到几百步）	可达数千 Token
模型规模	通常几百万参数	数十亿到数千亿参数
采样成本	低（模拟器快速运行）	极高（LLM 自回归生成很慢）

显存是最大的瓶颈

PPO-RLHF 需要在显存中同时维护 **4 个 LLM 规模的模型**：

模型	参数量级	是否需要梯度
Policy Model π_θ	1× LLM	✅ 需要
Critic Model V_ψ	1× LLM	✅ 需要
Reward Model r_φ	1× LLM	❌ 仅推理
Reference Model π_{ref}	1× LLM	❌ 仅推理

对于一个 7B 参数的模型，仅模型权重就需要约 **4 × 7B × 2 bytes (FP16) ≈ 56 GB** 显存，加上梯度、优化器状态和激活值，实际需求远超 80GB 单卡容量。

2.6 PPO 的痛点总结

痛点	具体表现
显存占用极高	需要 4 个 LLM 规模的模型同时在显存中
Critic 训练困难	在奖励稀疏的场景下，价值模型的准确估计本身就是难题
超参数敏感	clip 阈值 ε、KL 系数 β、学习率、GAE 参数等相互耦合，调参困难
训练不稳定	长序列生成 + 稀疏奖励 → 高方差梯度 → 容易崩溃
工程复杂	需要协调 4 个模型的训练、推理、通信，分布式实现难度大

关键问题：这些痛点中，哪些是可以通过算法设计来解决的？GRPO 的回答是：**Critic 可以去掉**。

唐国梁Tommy | TGLTommy 原创出品 · 侵权必究

唐国梁Tommy 原创出品｜© 2026 熵智未来（深圳）科技有限公司。保留所有权利。

本课程及其配套资料，包括但不限于课程视频、PDF 文档、课件、代码、图表、示例项目、知识库文档及其他学习材料，仅供购买本课程的学习者本人学习使用。除法律法规另有规定或经权利人书面许可外，不得复制、录屏、翻录、截图传播、转载、上传、分发、转售、改编、翻译、节选、汇编，或用于商业培训、企业内训、知识付费产品、AI 模型训练、微调、数据集构建、商业化知识库建设及其他商业用途。

本课程及配套资料中涉及的第三方开源软件、模型、代码、数据集、商标、图片或其他材料，相关权利归原权利人所有，并遵循其相应许可证或授权条款。

对于侵犯本课程及配套资料相关权益的行为，权利人将依法采取包括但不限于投诉、要求下架、要求停止侵害、索赔、追究法律责任等措施；构成违法犯罪的，将依法追究相应责任。

侵权举报 / 授权合作：*legal@tgltommyai.com*

🎯 **官方内容平台**：

- **微信公众号「唐国梁TGLTommy」**
- **B站「唐国梁Tommy」** / **「唐国梁Tommy的精品课」**
- **小红书「唐国梁TGLTommy」**
- **YouTube「唐国梁TGLTommy」**

🌐 **官方网站**：**TGLTommy.com**（国内访问需科学上网）