大语言模型的强化学习训练
提升推理、数学与代码能力
无需人类标注的自我学习方法

演讲者

技术分享

2025 年 12 月 5 日

目录

为什么需要强化学习训练 LLM?

传统方法的局限

- 监督微调 (SFT) 依赖大量人工标注数据
- 人类标注成本高、难以扩展
- 复杂推理任务的标注质量难以保证
- 模型只能学习到人类已知的解法

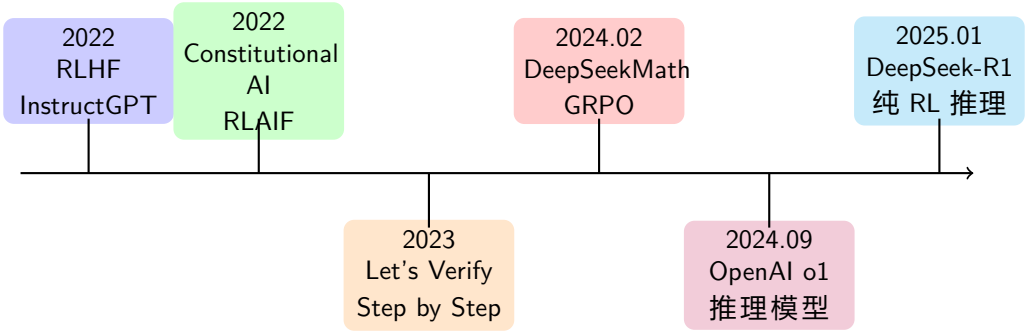
强化学习的优势

- 自动探索：模型自主发现解题策略
- 可扩展：无需人工标注
- 突破上限：可能发现人类未知的方法
- 自我改进：持续迭代优化

核心思想

通过**奖励信号**（如答案正确性、代码执行结果）引导模型自主学习推理能力

里程碑式进展



关键突破：DeepSeek-R1-Zero 首次证明无需任何 SFT 数据，仅通过 RL 即可获得强大推理能力

强化学习训练 LLM 的统一范式

梯度更新的统一形式

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{(q,o)\sim\mathcal{D}} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} \underbrace{GC(q, o, t, \pi_{rf})}_{\text{梯度系数}} \nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) \right]$$

方法	数据来源	奖励函数	训练方式
SFT	人工标注	-	离线
RFT	SFT 模型采样	规则	离线
DPO	SFT 模型采样	规则/偏好	离线
Online RFT	策略模型采样	规则	在线
PPO	策略模型采样	奖励模型	在线
GRPO	策略模型采样	奖励模型	在线

核心差异：数据来源 (在线/离线)、奖励信号、梯度系数计算方式

PPO 算法回顾

PPO 目标函数

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}_{q\sim P(Q),o\sim\pi_{\theta_{old}}}\left[\frac{1}{|o|}\sum_{t=1}^{|o|}\min\left(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t\right)\right]$$

其中：

- $r_t(\theta) = \frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})}$ 是重要性采样比率
- A_t 是优势函数，通过 GAE 计算：$A_t = \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}$
- ϵ 是裁剪参数（通常 0.1-0.2）

PPO 的问题：需要训练一个与策略模型同等规模的价值函数 (Critic)，显存开销大

GRPO 算法详解 (DeepSeek 核心创新)

Group Relative Policy Optimization 目标函数

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{q,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t}\left[\min(r_{i,t}\hat{A}_{i,t}, \text{clip}(r_{i,t})\hat{A}_{i,t}) - \beta\mathbb{D}_{KL}\right]$$

核心创新 - 组相对优势估计:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}$$

GRPO vs PPO

- 无需 Critic 模型
- 对同一问题采样 G 个回答
- 用组内相对奖励估计优势
- 显存节省约 50%

KL 散度正则化

$$\mathbb{D}_{KL} = \frac{\pi_{ref}(o_t|q, o_{<t})}{\pi_\theta(o_t|q, o_{<t})} - \log\frac{\pi_{ref}}{\pi_\theta} - 1$$

防止策略偏离参考模型太远

GRPO 算法伪代码

Algorithm 1 Iterative GRPO

Require: 初始策略 $\pi_{\theta_{init}}$, 奖励模型 r_ϕ, 提示集 $\mathcal{D}$

$\pi_\theta \leftarrow \pi_{\theta_{init}}$
for iteration $= 1, \ldots, I$ **do**
  $\pi_{ref} \leftarrow \pi_\theta$ {更新参考模型}
  for step $= 1, \ldots, M$ **do**
    从 $\mathcal{D}$ 采样一批提示 $\mathcal{D}_b$
    $\pi_{\theta_{old}} \leftarrow \pi_\theta$
    for 每个问题 $q \in \mathcal{D}_b$ **do**
      采样 G 个输出 $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)$
      计算奖励 $\{r_i\}_{i=1}^{G}$
      计算组相对优势 $\hat{A}_i = (r_i - \bar{r})/\sigma_r$
    end for
    通过最大化 $\mathcal{J}_{GRPO}$ 更新 π_θ
  end for
end for
return π_θ

奖励建模：结果监督 vs 过程监督

结果监督 (Outcome Supervision)

- 只在最后给出奖励
- 答案正确 $\rightarrow +1$
- 答案错误 $\rightarrow 0$ 或 -1
- 简单但信号稀疏

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$$

过程监督 (Process Supervision)

- 每个推理步骤都给奖励
- 需要过程奖励模型 (PRM)
- 信号密集，学习更高效
- 但 PRM 训练困难

$$\hat{A}_{i,t} = \sum_{index(j) \geq t} \tilde{r}_i^{index(j)}$$

OpenAI 研究结论 (Let's Verify Step by Step)

过程监督在 MATH 数据集上**显著优于**结果监督，解决率从约 70% 提升到 **78%**

DeepSeek-R1-Zero：纯 RL 的奇迹

训练设置

- 基座模型：DeepSeek-V3-Base
- 完全无 SFT 数据
- 仅使用规则奖励（答案正确性）
- 简单模板约束输出格式

涌现的能力

- 自我验证 (Self-verification)
- 反思 (Reflection)
- 长链思维 (Long CoT)
- "Aha Moment" - 顿悟时刻

模型	AIME 2024
GPT-4o	9.3%
Claude-3.5	16.0%
o1-mini	63.6%
o1-0912	74.4%
R1-Zero	**71.0%**
R1-Zero (maj)	86.7%

无需任何人类标注数据，达到 o1-0912 水平！

DeepSeek-R1-Zero 的" 顿悟时刻"

训练过程中观察到的现象

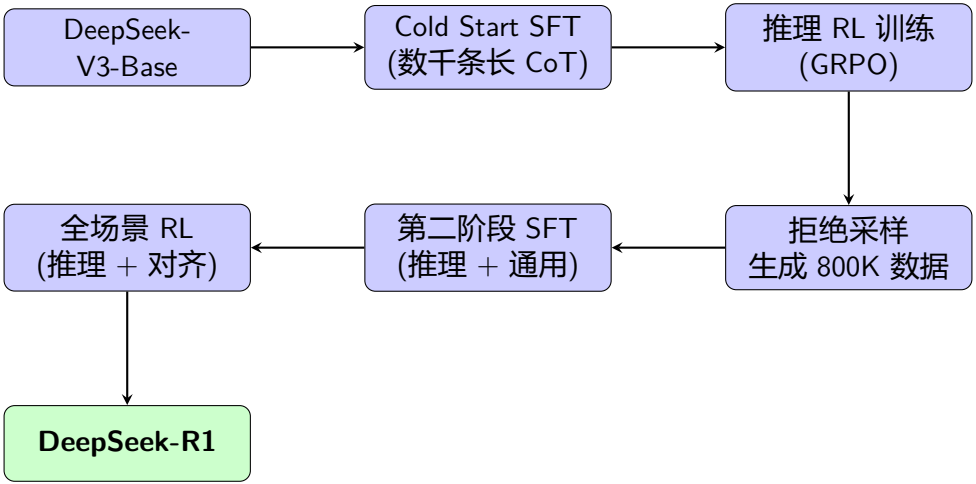
模型自主学会了重新评估和反思：

"Wait, let me reconsider. I made an error in my calculation..."
"Hmm, but wait, let me re-examine this..."
"I realize there's another approach that might work better..."

关键发现：我们没有显式教模型如何解决问题，
只是给了它正确的激励，它就自
主发展出了高级的问题解决策略。
这展示了强化学习释放 AI 智能的潜力。

思考时间自动增长：随着训练，模型输出长度从几百 token 增长到数千 token

DeepSeek-R1 完整训练流程



四阶段训练:

1. **Cold Start**: 少量高质量长 CoT 数据启动
2. **推理 RL**: 专注数学、代码、逻辑推理

DeepSeek-R1 性能对比

模型	AIME	MATH	GPQA	LiveCode	MMLU	CF Rating
GPT-4o	9.3	74.6	49.9	32.9	87.2	759
Claude-3.5	16.0	78.3	65.0	38.9	88.3	717
DeepSeek-V3	39.2	90.2	59.1	36.2	88.5	1134
o1-mini	63.6	90.0	60.0	53.8	85.2	1820
o1-1217	79.2	96.4	75.7	63.4	91.8	2061
DeepSeek-R1	**79.8**	**97.3**	71.5	**65.9**	90.8	**2029**

- AIME 2024：79.8%，超越 o1-1217
- MATH-500：97.3%，与 o1-1217 持平
- Codeforces：2029 Elo，超过 96.3% 人类选手
- 首个开源达到 o1 水平的推理模型

知识蒸馏：让小模型也能推理

蒸馏方法

- 使用 DeepSeek-R1 生成 800K 训练样本
- 直接 SFT 微调小模型
- 无需额外 RL 训练

关键发现

- 蒸馏 $>$ 小模型直接 RL
- 大模型发现的推理模式更优
- 32B 蒸馏模型超越 QwQ-32B

模型	AIME	MATH
QwQ-32B	50.0	90.6
R1-Distill-1.5B	28.9	83.9
R1-Distill-7B	55.5	92.8
R1-Distill-14B	69.7	93.9
R1-Distill-32B	72.6	94.3
R1-Distill-70B	70.0	94.5

7B 蒸馏模型超越 GPT-4o!

OpenAI o1：官方披露的信息

已知信息

- 使用大规模强化学习训练
- Chain-of-Thought 推理
- 测试时计算可扩展
- 隐藏思维链（安全考虑）

性能数据

- AIME 2024: 74% → 83% (maj@64) → 93%
- IOI 2024: 213 分 (49th percentile)
- 放宽限制后: 362 分 (金牌水平)

关键声明

"Our large-scale RL algorithm teaches the model how to think productively using its chain of thought in a highly **data-efficient** training process."

性能随训练计算和测试计算同时提升

OpenAI o1 技术推测

基于公开信息和研究趋势的推测

1. **训练方法**
   - 可能使用 PPO 或其变体进行大规模 RL
   - 结合过程奖励模型 (PRM) 和结果奖励
   - 使用专家标注的步骤级反馈数据 (如 PRM800K)
2. **推理增强**
   - 训练模型生成长链思维
   - 可能使用搜索算法 (Beam Search/MCTS) 优化输出
   - 学习的评分函数用于重排候选答案
3. **安全对齐**
   - 将安全规则融入思维链
   - 思维链提供可解释性和监控能力
   - 不直接展示原始思维链给用户

Google 的推理增强方法

Minerva (2022)

- 基于 PaLM，在数学数据上继续预训练
- 540B 参数，MATH 达到 33.6%
- 主要依赖预训练而非 RL

AlphaGeometry (2024)

- 专门解决几何问题
- 结合神经网络和符号推理
- 不依赖人类示范

Gemini 系列

- Gemini Pro/Ultra 具备强推理能力
- 训练细节未公开
- 推测使用了类似 RLHF 的方法

Anthropic: Constitutional AI (RLAIF)

核心思想：用 AI 反馈代替人类反馈

两阶段训练：

1. **监督学习阶段**
   - 模型生成回答
   - 根据" 宪法" 原则自我批评
   - 生成修订后的回答
   - 在修订回答上微调
2. **RL 阶段 (RLAIF)**
   - 生成多个回答
   - 用 AI 模型评估哪个更好
   - 训练偏好模型
   - 使用偏好模型做 RL

优势：大大减少人类标注需求，可扩展性强

奖励设计的关键考虑

1. **规则奖励 vs 模型奖励**
   - **规则奖励**：答案正确性、代码执行结果
     - 优点：准确、无 reward hacking
     - 缺点：只适用于有标准答案的任务
   - **模型奖励**：训练奖励模型评分
     - 优点：可用于开放式任务
     - 缺点：易被 hacking，需要迭代更新

2. **DeepSeek-R1 的选择**
   - R1-Zero：纯规则奖励（答案正确性 + 格式）
   - 避免神经奖励模型带来的 reward hacking
   - 简化训练流程，无需重复训练奖励模型

训练超参数与技巧

参数	DeepSeekMath GRPO 设置
策略模型学习率	1e-6
KL 系数 β	0.04
每问题采样数 G	64
最大输出长度	1024 tokens
训练批大小	1024
裁剪参数 ϵ	0.2 (典型值)

关键训练技巧:

- 每次探索后只更新一次策略（稳定性）
- 使用 KL 散度约束防止策略漂移
- 迭代 RL：定期更新参考模型和奖励模型
- 语言一致性奖励：防止语言混杂

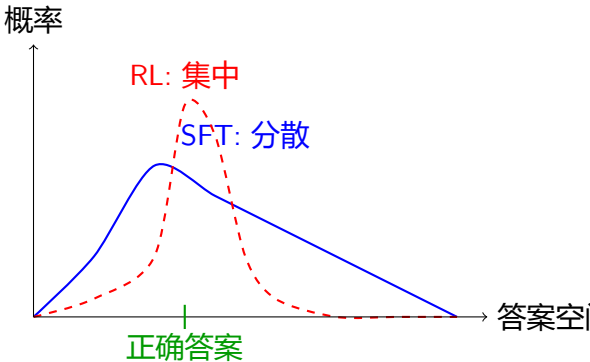
为什么 RL 有效？深入分析

实验发现

- RL 提升 Maj@K 但不提升 Pass@K
- 说明 RL 主要稳定输出分布
- 而非增加基础能力

Maj@K vs Pass@K

- **Pass@K**：K 个样本中至少一个正确
- **Maj@K**：K 个样本多数投票结果
- RL 让正确答案更容易被采样到



RL 使输出分布更集中于正确答案

DeepSeek 团队的失败尝试

1. **过程奖励模型 (PRM)**
   - 难以定义通用推理的" 步骤"
   - 中间步骤正确性难以自动判断
   - 模型 PRM 容易被 hacking
   - 重新训练 PRM 增加复杂度

2. **蒙特卡洛树搜索 (MCTS)**
   - Token 生成的搜索空间指数级大
   - 价值模型训练困难
   - 难以像 AlphaGo 那样迭代提升
   - 推理时有效，但训练时难以自我改进

启示：简单的方法 (GRPO+ 规则奖励) 反而更有效、更可扩展

未来研究方向

1. **数据层面**
   - 使用分布外问题进行探索
   - 高级采样策略（如树搜索）
   - 高效推理加速探索

2. **算法层面**
   - 对噪声奖励信号的鲁棒算法
   - Weak-to-Strong 学习
   - 更好的过程监督方法

3. **奖励函数**
   - 提高奖励模型泛化能力
   - 建模奖励的不确定性
   - 高效构建高质量过程奖励模型

核心要点总结

技术突破

- **GRPO**：无需 Critic 的高效 RL 算法
- **规则奖励**：简单有效，避免 reward hacking
- **纯 RL 训练**：无需 SFT 即可涌现推理能力
- **知识蒸馏**：大模型推理能力可迁移到小模型

关键洞见

- RL 主要通过稳定输出分布提升性能
- 简单方法往往比复杂方法更有效
- 正确的激励机制可以让 AI 自主发展高级能力
- 测试时计算和训练时计算都很重要

主流模型训练方法对比

	DeepSeek-R1	OpenAI o1	Claude	Gemini
基础方法	GRPO	PPO 变体 (推测)	RLAIF	未公开
是否需要 SFT	可选	未知	是	未知
奖励类型	规则为主	PRM+ORM(推测)	AI 反馈	未知
开源	✓	×	×	×
训练细节	公开	保密	部分	保密

DeepSeek-R1 的意义:
首次开源证明纯 RL 可以达到顶级推理能力,
为社区提供了可复现的研究基础

参考文献

- DeepSeek-AI. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via RL*. arXiv:2501.12948, 2025.
- Shao et al. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning*. arXiv:2402.03300, 2024.
- OpenAI. *Learning to Reason with LLMs*. 2024.
- Lightman et al. *Let's Verify Step by Step*. arXiv:2305.20050, 2023.
- Bai et al. *Constitutional AI: Harmlessness from AI Feedback*. arXiv:2212.08073, 2022.
- Schulman et al. *Proximal Policy Optimization Algorithms*. arXiv:1707.06347, 2017.
- Luo et al. *WizardMath: Empowering Mathematical Reasoning via Reinforced Evol-Instruct*. ICLR 2025.

谢谢!

Q&A