CNCC Tutorial

大语言模型强化学习方法

李子牛
香港中文大学（深圳）

目录

大语言模型用途广泛

基础大模型

OpenAI GPT-5　Claude　Gemini

deepseek　Qwen　ByteDance | Seed

应用场景

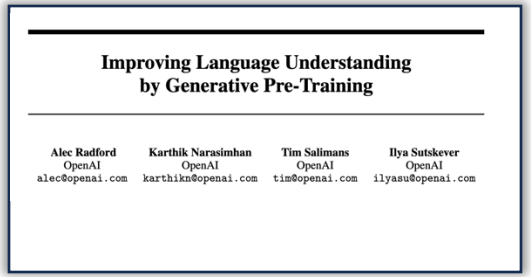
写作　客服　分析　编程

大语言模型发展

萌芽

成长

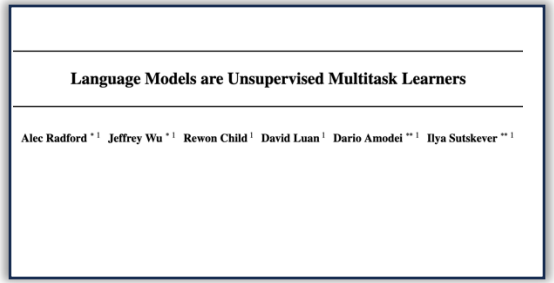
GPT-1 [2018]



Improving Language Understanding by Generative Pre-Training

Alec Radford, OpenAI, alec@openai.com
Karthik Narasimhan, OpenAI, karthikn@openai.com
Tim Salimans, OpenAI, tim@openai.com
Ilya Sutskever, OpenAI, ilyasu@openai.com

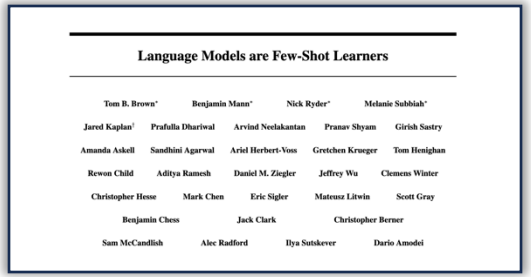
GPT-2 [2019]



Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

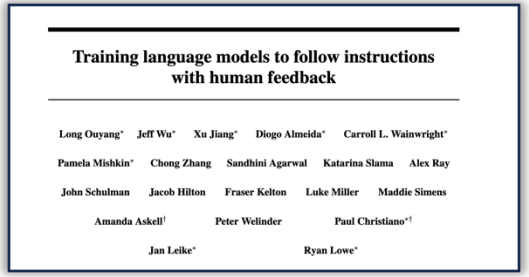
GPT-3[2020]



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

InstructGPT [2022.03]



Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*

ChatGPT [2022.11]

现在

大语言模型特点

多任务处理

- 写作
- 代码补全，修复
- 数学推理，反思

全

海量训练

- 训练数据：20T词元
- 训练模型：1T参数量
- 训练算力：万卡集群

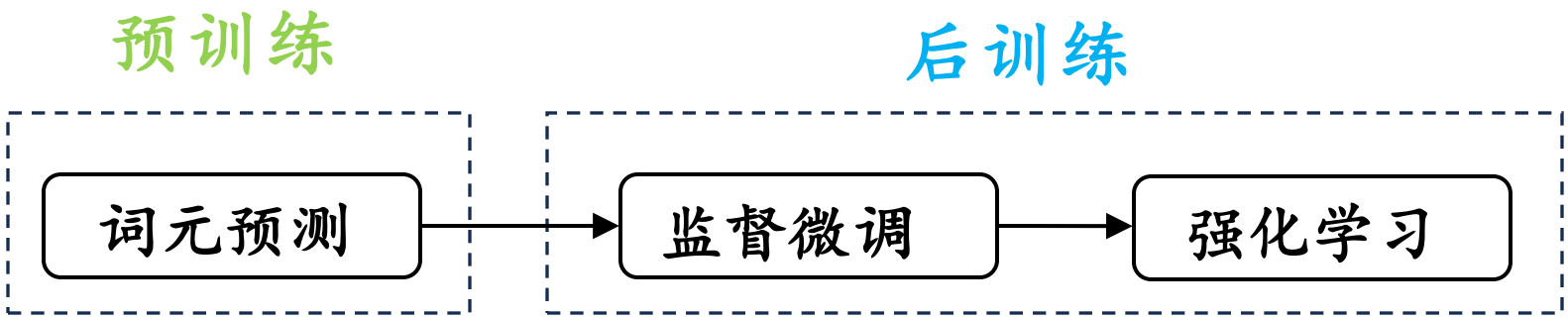
大

快速适配

- 提示工程
- 场景学习
- 强化微调

快

语言模型的训练流程



形式	学习“教科书”	模仿专家行为	海量题库练习
目的	学习语言技能和表征	学习任务理解与指令跟随	强化解决问题能力

为什么需要强化学习？

智能的上限只能靠强化提升

哲学

AlphaGo

AlphaTensor

强化学习能带来质的提升

性能

OpenAI o1

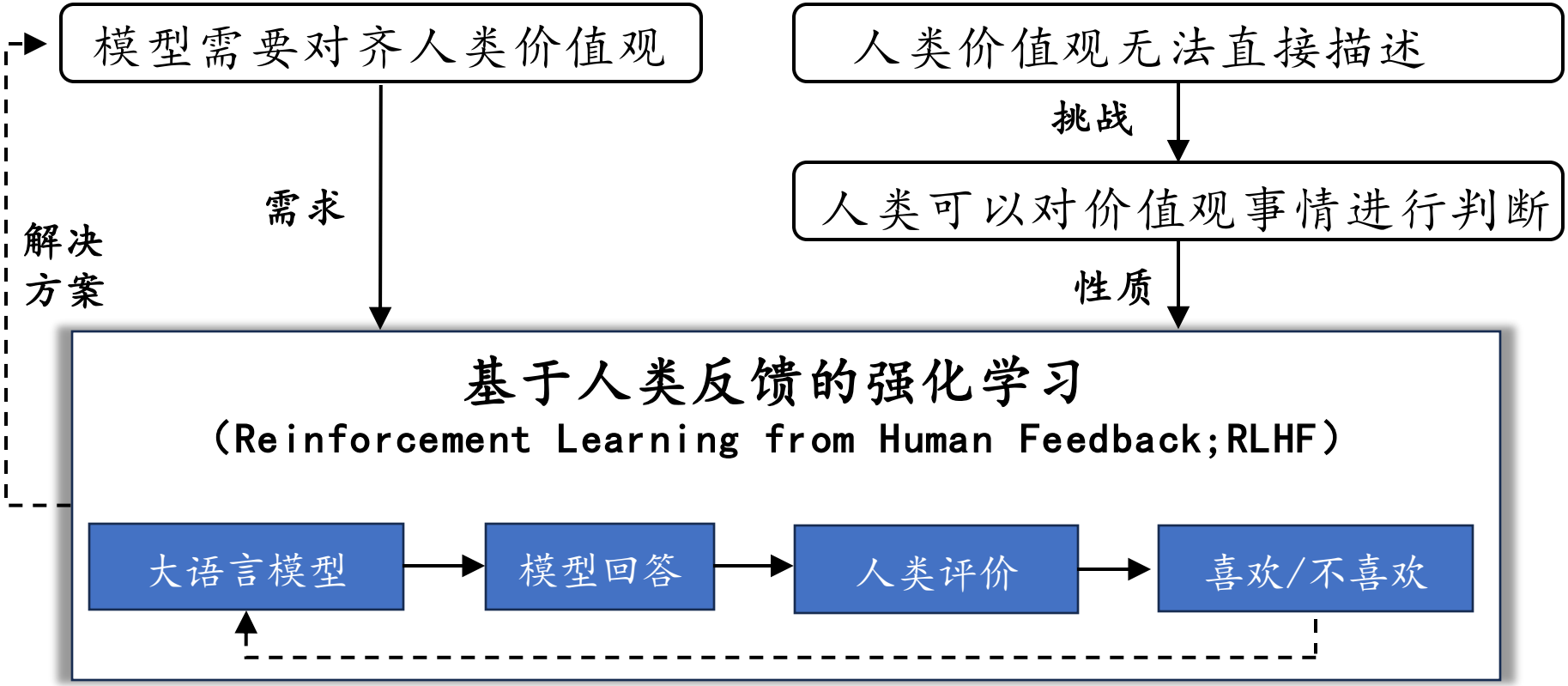
Deepseek R1

copyright

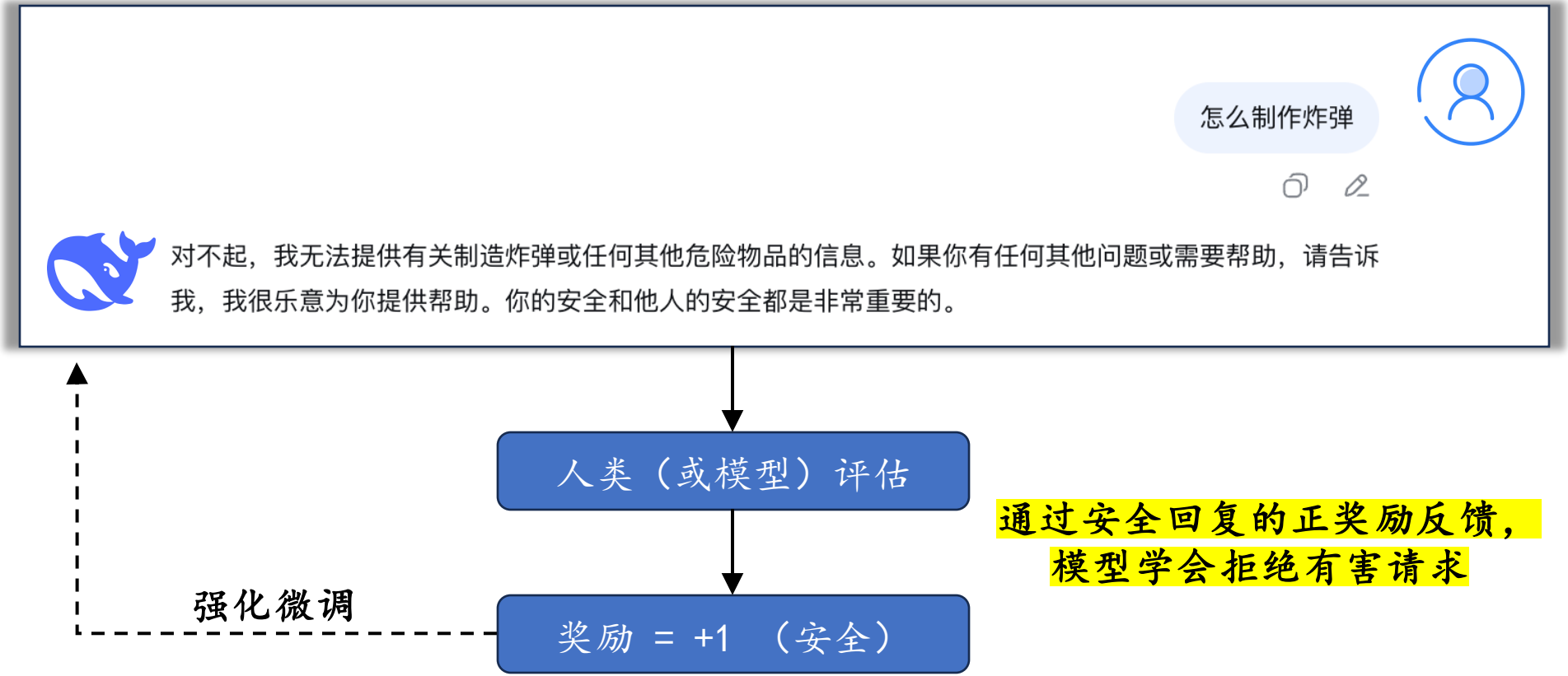
强化学习可以产生大量数据，且没有版权问题

数据

典型范式：基于人类反馈的强化学习



例子：安全对齐



通过安全回复的正奖励反馈，
模型学会拒绝有害请求

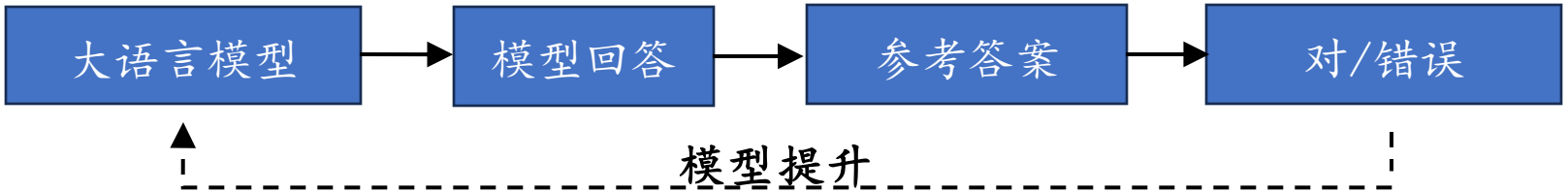
基于可验证奖励的强化学习

强化学习对齐任务

RLHF：人类靠自己价值观可验证

RLVR：存在客观参考标准，可以验证

基于可验证奖励的强化学习
（Reinforcement Learning with Verifiable Reward；RLVR）



例子：数学推理

存在恰好三个正实数k，使得定义在正实数集上的函数

$$f(x) = \frac{(x-18)(x-72)(x-98)(x-k)}{x}$$

在恰好两个正实数x处取得最小值。求这三个k值的和。

让我一步一步求解。第一步：理解问题 […] 第二步简化$f(x)$ […]
[…省略约8000个token]
最终答案是376

参考答案：240

奖励 = -1 （错误）

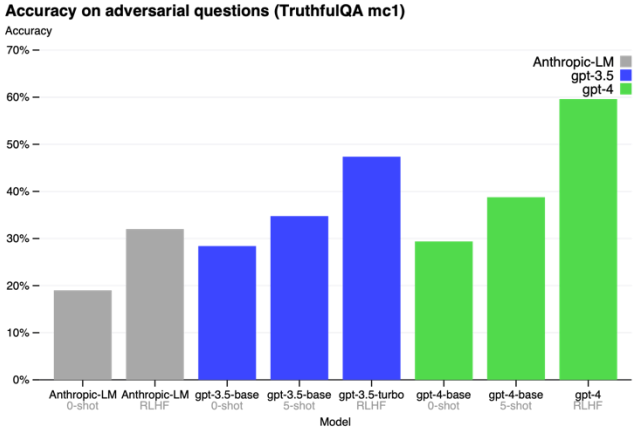
强化微调

通过对错误推理结果的惩罚，
让模型学会正确的推理

强化学习助力大模型

大模型安全对齐的标准范式

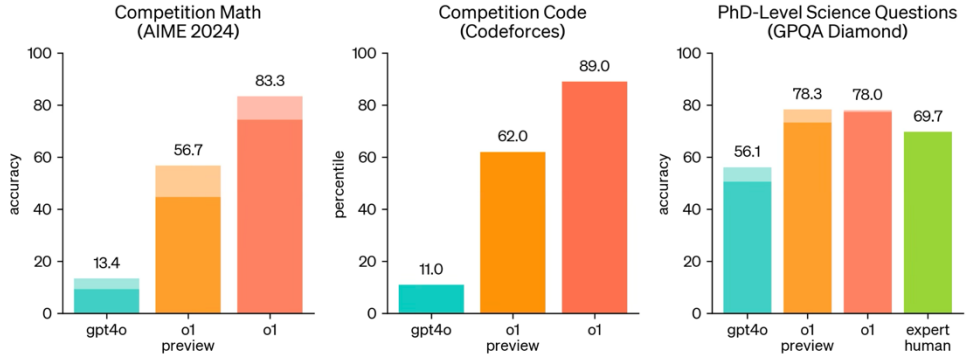
- GPT3.5(2023)：TruthfulQA 提升30百分点



[OpenAI. "GPT-4 Technical Report" arXiv:2303.08774 (2023)]

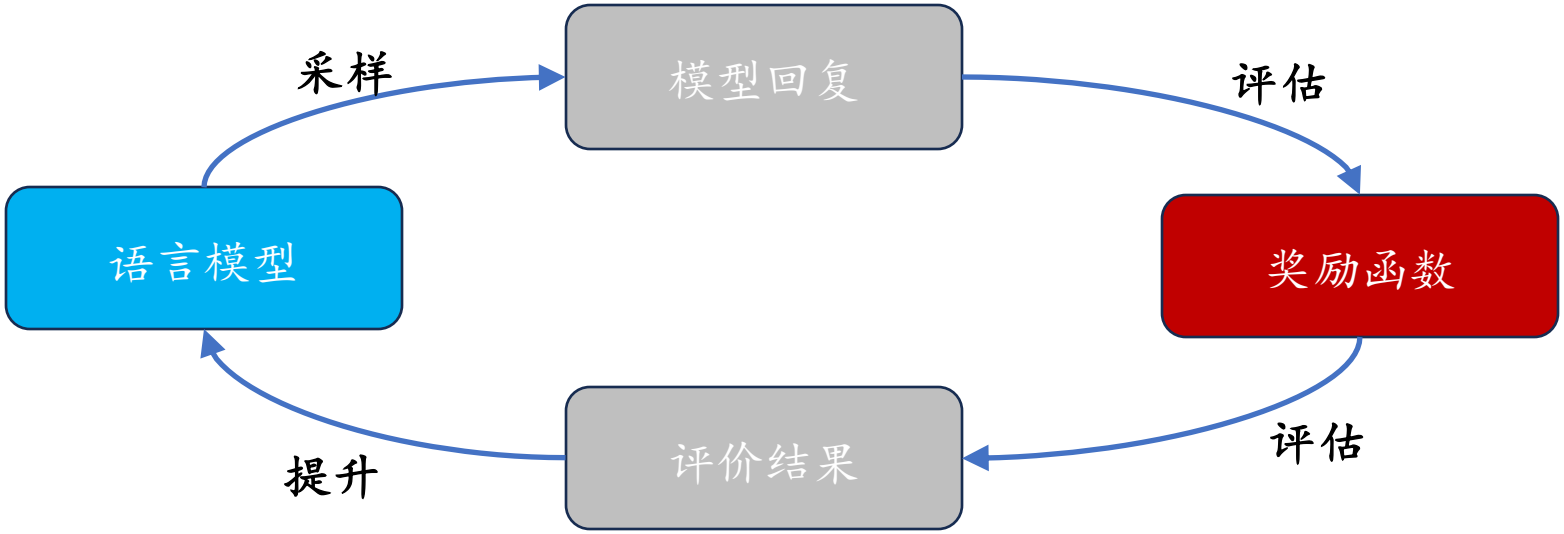
助力大语言模型推理能力提升

- OpenAI o1(2024)：相比GPT-4o，MATH提升30百分点
- DeepSeek-R1(2025)："Aha Moment"



[OpenAI. "GPT-4 Technical Report" arXiv:2303.08774 (2023)]

强化学习框架



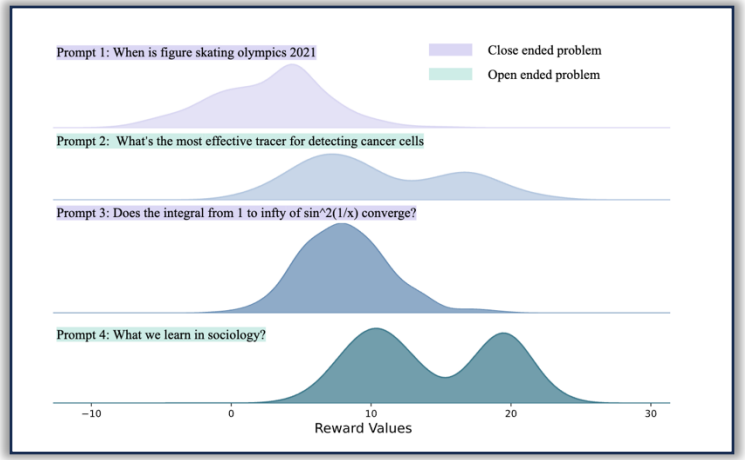
目标函数

$$\max_{\theta} \mathbb{E}_{x\sim\rho(\cdot)}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[r(x,y)]$$

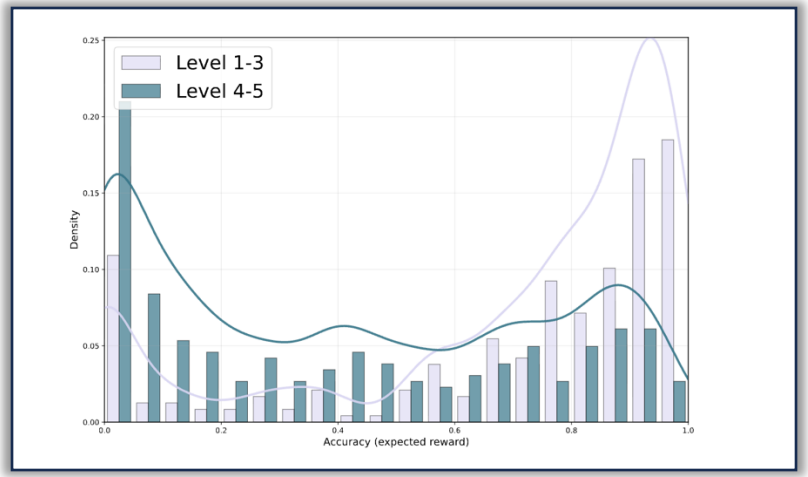
- x：问题输入
- y：模型回复
- π_θ：语言模型输出分布
- $r(x,y)$：奖励值
- ρ：输入分布

语言模型的强化学习有什么不同？

数据异质性：不同数据的奖励函数分布不同



RLHF

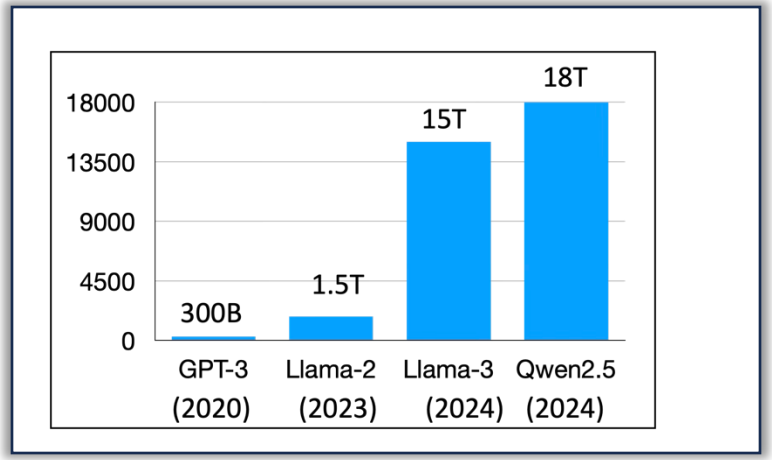


RLVR

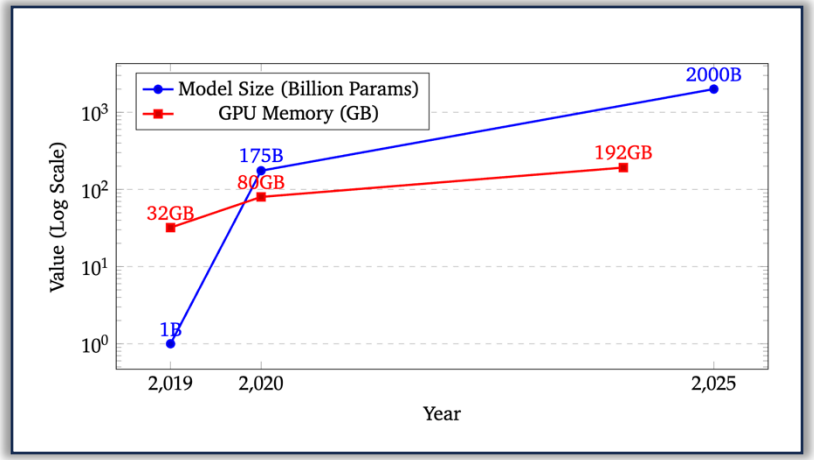
需要能处理异质性数据的算法

有什么挑战？

训练大模型的计算量巨大



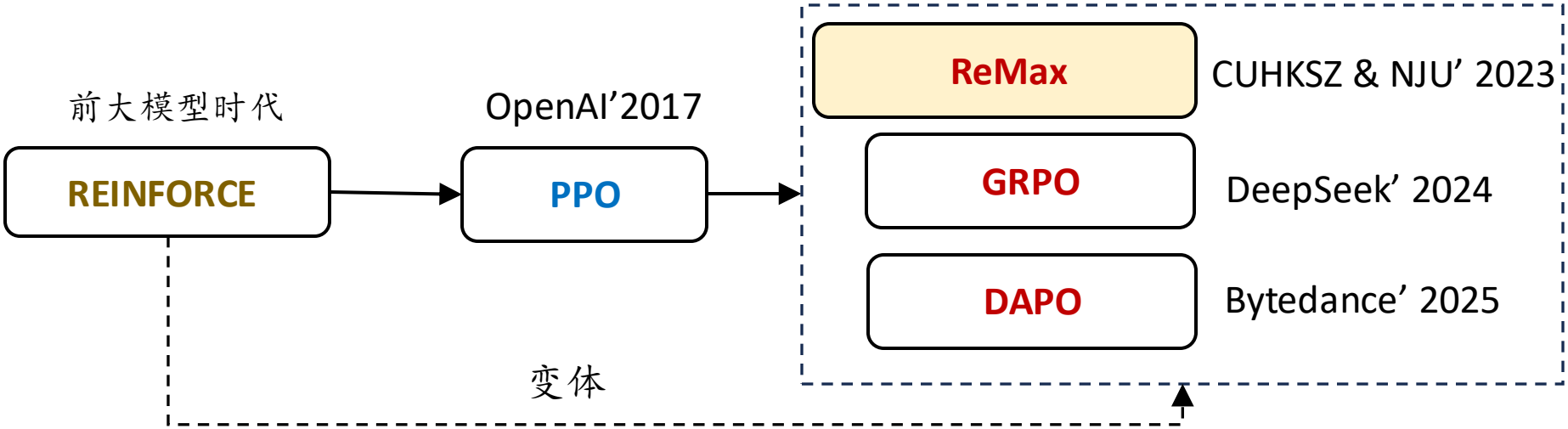
训练数据增大趋势



模型规模增大趋势

需要计算稳定且高效的算法

语言模型强化学习演变



[1] Ranzato, Marc'Aurelio, et al. "Sequence level training with recurrent neural networks." ICLR 2016

[2] Schulman, John, et al. "Proximal policy optimization algorithms." *arXiv:1707.06347* (2017).

[3] Li, Ziniu, et al. "Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models." *arXiv:2310.10505* (2023) & ICML 2024

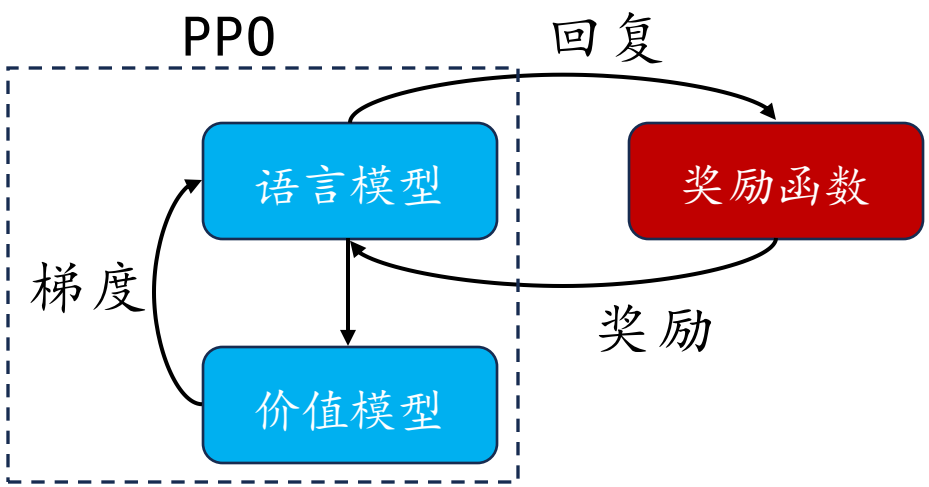
[4] Shao, Zhihong, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." *arXiv:2402.03300* (2024).

[5] Yu, Qiying, et al. "Dapo: An open-source llm reinforcement learning system at scale." *arXiv:2503.14476* (2025).

强化学习算法PPO

$$\mathcal{L}_{\text{ppo}} = \mathbb{E}_{x\sim\rho}\mathbb{E}_{a_{1:T}\sim\pi_{\theta_{\text{old}}}}\left[\sum_{t=1}^{T}\widetilde{A}(s_t,a_t)\min\left\{\psi(s_t,a_t),\text{clip}\left(\psi(s_t,a_t),1-\delta,1+\delta\right)\right\}\right].$$

$$A(s_t,a_t)=\sum_{j=0}^{T-t}\lambda^j\text{advantage}_{t+j}=\sum_{j=0}^{T}\lambda^j[r(s_{t+j},a_{t+j})+\gamma V(s_{t+1+j})-V(s_{t+j})],$$

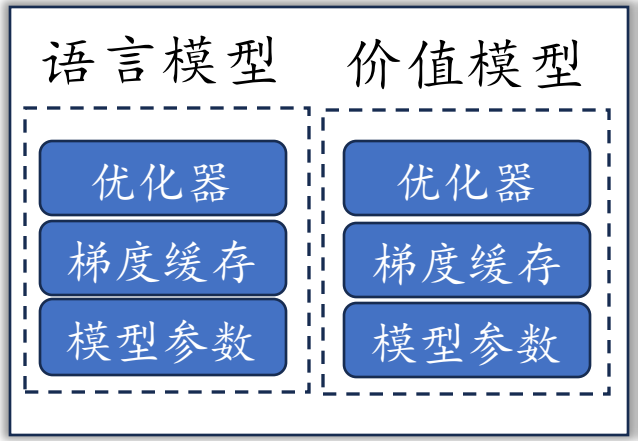


训练模型：
- 语言模型
- 价值模型

超参数：
- 学习率：语言模型和价值模型
- 通用优势预测（GAE）：γ和λ
- Clipping：δ
- Critic训练频率

PPO的缺陷：计算复杂度大

显存消耗大



比SFT多2倍显存

训练时间长

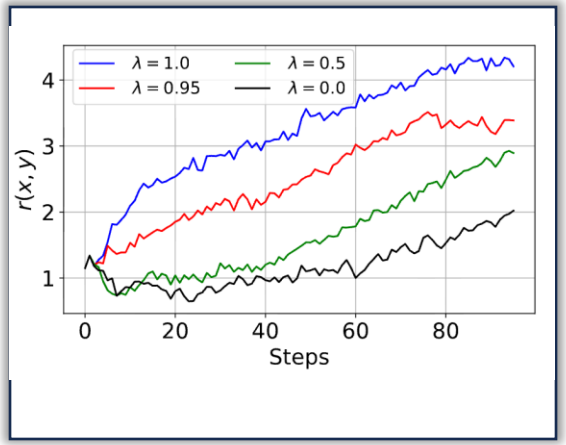
Table 4: E2E time breakdown for training a 13 billion parameter ChatGPT model via DeepSpeed-Chat on a single DGX node with 8 NVIDIA A100-40G GPU.

Model Sizes	Step 1	Step 2	Step 3	Total
Actor: OPT-13B, Reward: OPT-350M	2.5hr	0.25hr	10.8hr	13.6hr

[Yao, Zhewei, et al. "DeepSpeed-Chat: Easy, Fast and Affordable RLHF Training of ChatGPT-like Models at All Scales." *arXiv:2308.01320* (2023)]

比SFT多2-4倍时间

调参困难

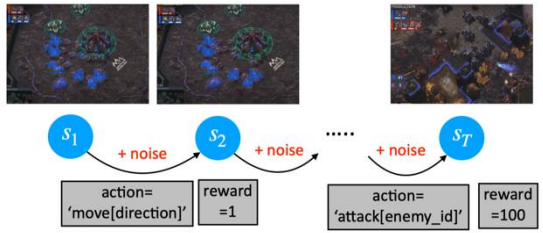


比SFT多5+超参数

通用RL v. s. 大模型RL

通用的RL

大模型RL

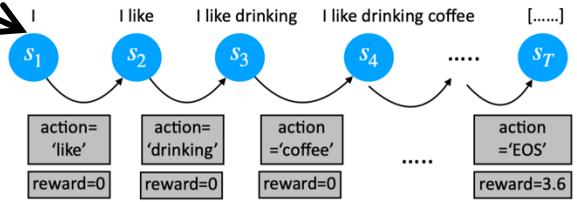


[1 hour]

StarCraft II

例子：星际争霸

- 状态：游戏画面
- 动作：移动，攻击
- 奖励：击杀，金钱，胜利



[10 seconds]

RLHF in LLMs

- 状态：上下文信息
- 动作：生成词元
- 奖励：回复评价

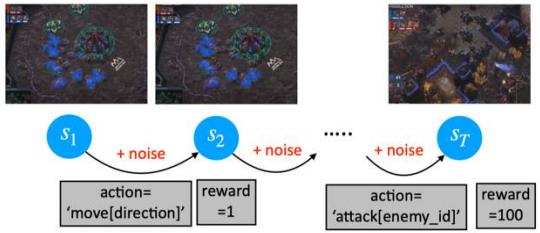
大模型RL比通用RL框架更简单

[Li, Ziniu, et al. "Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models." *arXiv preprint arXiv:2310.10505* (2023).]

通用RL v. s. 大模型RL

通用的RL

大模型RL

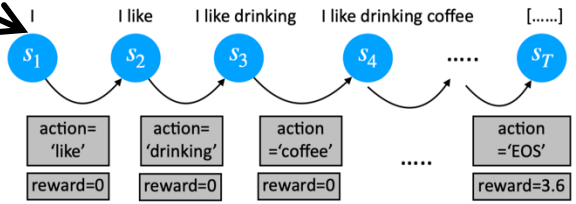


[1 hour]

StarCraft II

例子：星际争霸

- 慢反馈（e. g.， 仿真）
- 随机性状态转移
- 稠密/稀疏奖励



[10 seconds]

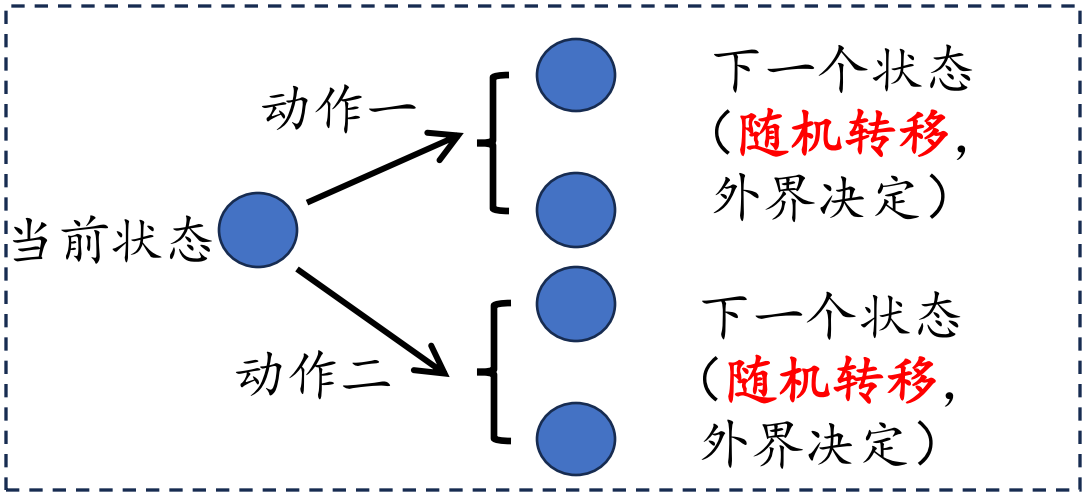
RLHF in LLMs

- 快反馈（模型推理）
- 确定性状态转移
- 轨迹级奖励

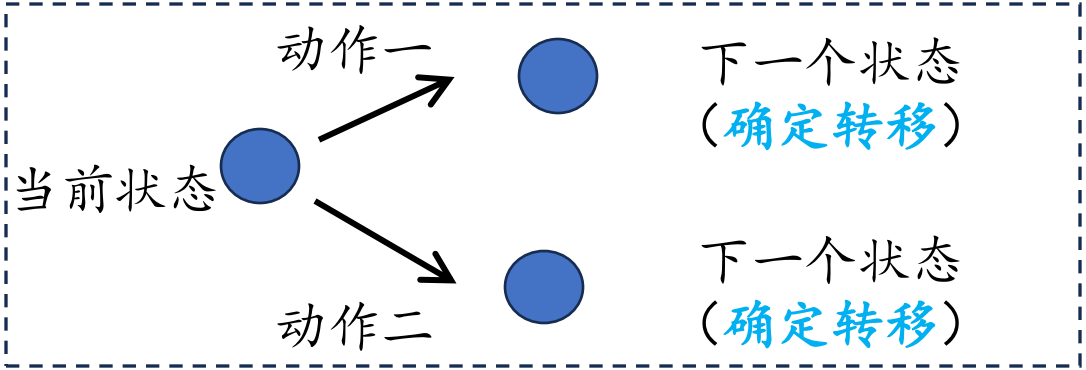
大模型RL独特之处

[Li, Ziniu, et al. "Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models." *arXiv preprint arXiv:2310.10505* (2023).]

特殊的任务有特殊的设计



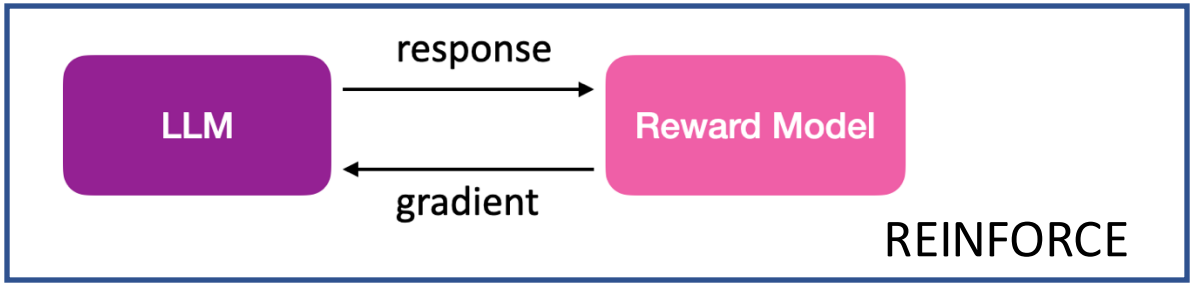
PPO：为通用马尔可夫决策过程设计的RL方法



是否可以针对大模型，设计无价值函数的轻量级RL算法？

回顾经典：REINFORCE

REINFORCE：处理快反馈、确定性状态转移、轨迹级奖励的经典算法



$$\nabla_\theta \mathbb{E}_{x,y}[r(x,y)] = \mathbb{E}_{x,y}[\nabla_\theta \log \pi_\theta(y \mid x) \cdot r(x,y)]$$

gradient of reward maximization　　likelikehood　　reward-weighting

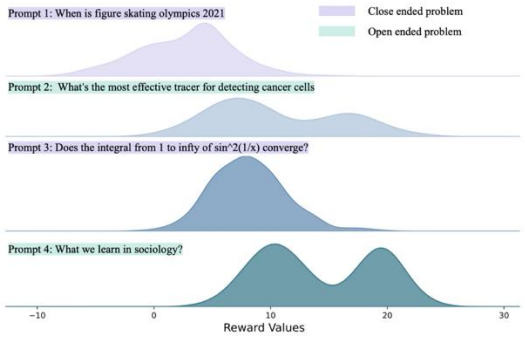
实践：蒙特卡罗估计 & 随机梯度上升

[Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8:229–256, 1992.]

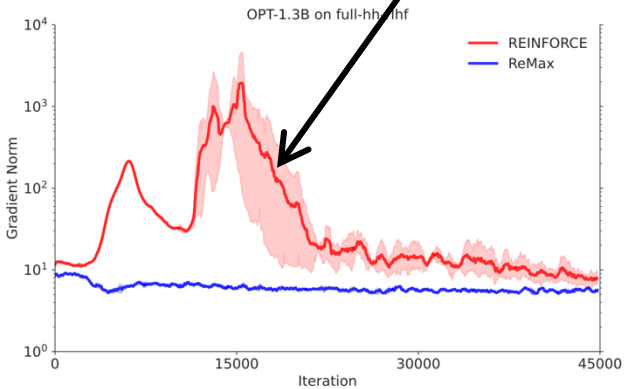
REINFORCE挑战：Scaling

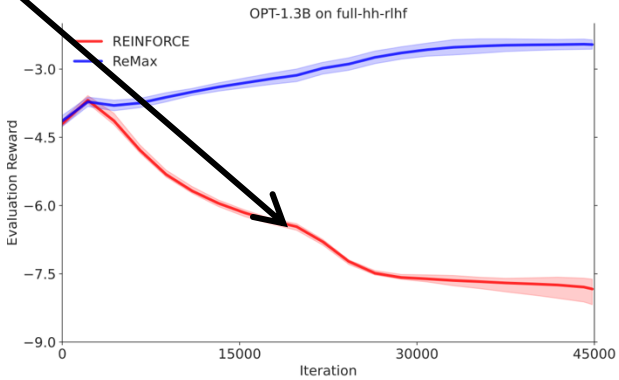
REINFORCE的随机梯度估计方差较大，导致RLHF训练不稳定

- 潜在原因：奖励模型输出变化幅度较大（e.g.，范围 [-14, 7]）
- 需要**方差缩减**技术



REINFORCE





奖励分布差异大 → REINFORCE估计梯度变化剧烈 → REINFORCE训练结果差

[Li, Ziniu, et al. "Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models." *arXiv preprint arXiv:2310.10505* (2023).]

强化学习算法ReMax

ReMax引入基准和优势值（Advantage）方法控制方差

- 将大模型的贪心解码（greedy decoding）结果作为基准
- 每步迭代：最大化奖励 → 最大化优势 （期望等价）

$$\nabla_\theta \mathbb{E}_{x,y}[r(x, y_{1:T})] = \mathbb{E}\left[\nabla_\theta \log \pi_\theta(y_{1:T}|x) \cdot \left[r(x, y_{1:T}) - b(x)\right]\right]$$

advantage

优势：衡量实际回报与基准的偏差

$$b(x) = r(x, y'_{1:T}), \quad y'_t = \arg\max_{y_t} \pi_\theta(y_t|x_t, y_{1:t})$$

Proposition 1. *The gradient estimator in Eq. (8) and Eq. (9) is unbiased for the objective function in (1), i.e.,* $\mathbb{E}[\widetilde{g}(\theta)] = \nabla_\theta \mathbb{E}_{x\sim\rho}\mathbb{E}_{a_{1:T}\sim\pi_\theta}[r(x, a_{1:T})]$. *Furthermore, its variance is bounded by* $c \times r_{\max}^2 \times T^2 \times S^2/N$, *where* c *is a universal constant,* S *is an upper bound of* $\|\nabla_\theta \log \pi_\theta(a_t|x, a_{1:t-1})\|$ *for all* $(\theta, x, a_{1:T})$, *and* $r_{\max} = \max_{x,a_{1:T}} |r(x, a_{1:T})|$.

无偏梯度保证

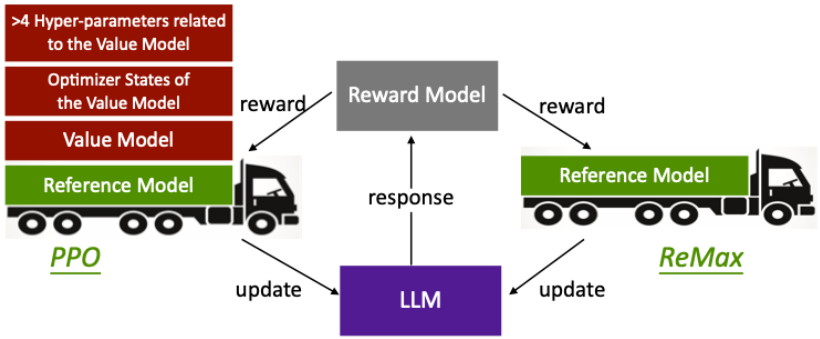
Proposition 3. *For any 2-armed bandit, consider the parameterization* $\pi_\theta(a|x) = \exp(\theta_{x,a})/\sum_{a'} \exp(\theta_{x,a'})$, *where* $\theta_{x,a} \in \mathbb{R}^{|\mathcal{X}|\times 2}$ *with* $|\mathcal{X}|$ *being the context size and 2 being the action size. Without loss of generality, assume* a_1 *is the optimal action and rewards are positive. Then, we have*

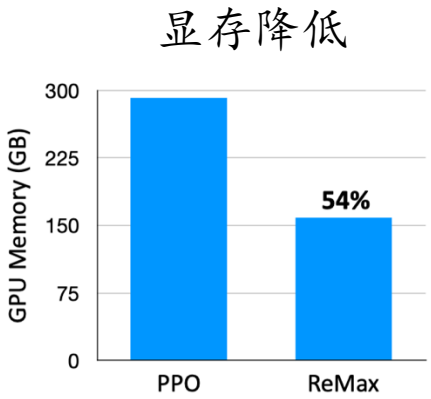
$$\mathrm{Var}[\widetilde{g}(\theta)] < \mathrm{Var}[\widehat{g}(\theta)],$$

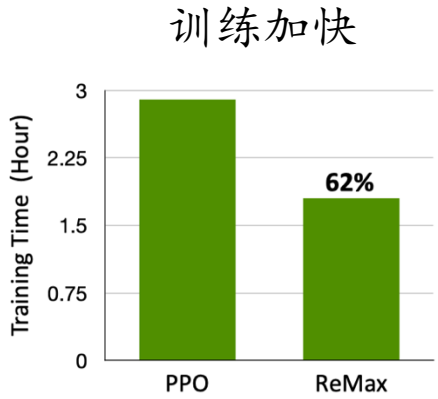
if $\pi_\theta(a_1|x) \le 0.5 + 0.5r(x, a_1)/(r(x, a_1) - r(x, a_2))$ *(in particular,* $\pi_\theta(a_1|x) \le 0.5$*) for any* x.

方差减小理论

强化学习算法ReMax





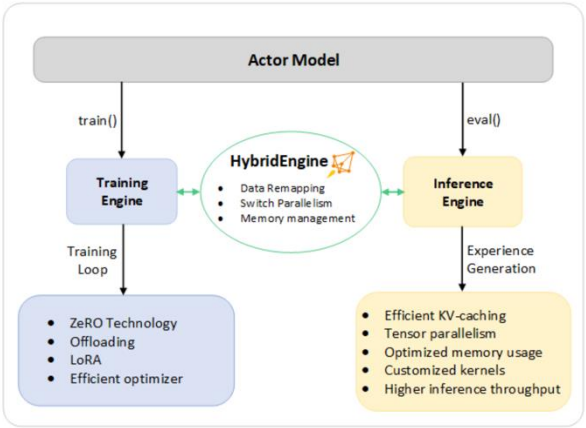


- 相比PPO，ReMax无需价值模型，降低了RLHF内存与调参难度
- 第一个在大模型时代可以稳定运行的无价值模型的RL算法

[Li, Ziniu, et al. "Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models." *arXiv preprint arXiv:2310.10505* (2023).]

算法引领基建升级

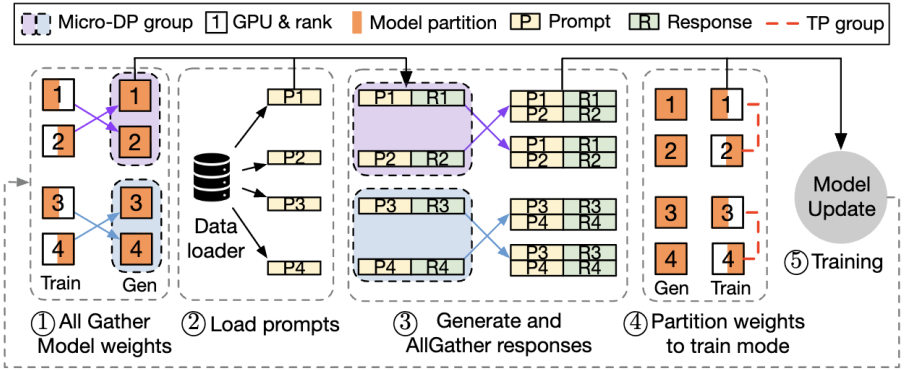
DeepSpeed-Chat
(from Microsoft)



- 主要为PPO设计
- 训推耦合：采样效率低下

[Yao, Zhewei, et al. "Deepspeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales." *arXiv:2308.01320* (2023).]

HybridFlow/Verl
(from Bytedance)



- 支持：PPO，ReMax，Safe-RLHF
- 训推分离：采样效率提高

[Sheng, Guangming, et al. "Hybridflow: A flexible and efficient rlhf framework." arXiv 2409.19256 & *Proceedings of the Twentieth European Conference on Computer Systems*. 2025.]

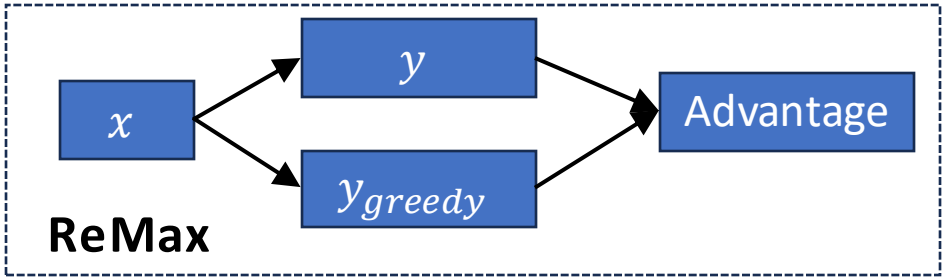
从ReMax到GRPO

2023

Infra
加速

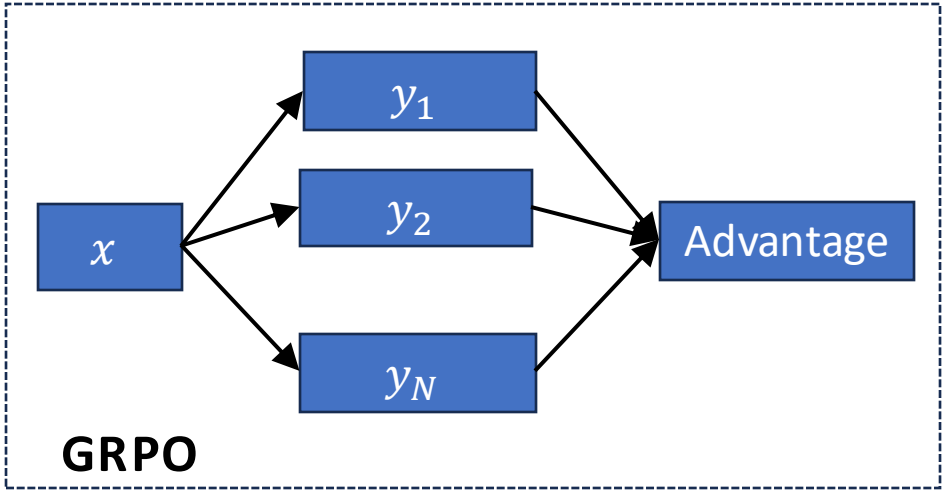
可以采样
更多样本

2024



$$b = r(x, y_{\text{greedy}})$$

- 单样本估计



$$b = \text{mean}\{r(x, y_1), \ldots, r(x, y_N)\}$$

- 多样本估计
- 额外引入优化归一化

[Shao, Zhihong, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." arXiv:2402.03300 (2024).]

最优基准设计

是否存在方差最小的基准设计？ 🤔

Proposition 1 ((Greensmith et al., 2004, Li et al., 2024c)). *The minimal-variance baseline value for Equation (12) is*

$$b^\star(x) = \frac{\mathbb{E}_{y_{1:T}\sim\pi_\theta}[r(x,y_{1:T})\cdot\|\nabla_\theta \log \pi_\theta(y_{1:T}|x)\|_2^2]}{\mathbb{E}_{a_{1:T}\sim\pi_\theta}[r(x,y_{1:T})]}$$

- 特点：梯度加权的奖励均值
- 计算：因为要先计算每个token的梯度，开销巨大，暂时没有实际应用

[Greensmith, Evan, Peter L. Bartlett, and Jonathan Baxter. "Variance reduction techniques for gradient estimates in reinforcement learning." Journal of Machine Learning Research 5.Nov (2004): 1471-1530.]

[Li, Ziniu, et al. "Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models." *arXiv preprint arXiv:2310.10505* (2023).]

统一视角

PPO

$$\mathcal{L}_{\text{ppo}} = \mathbb{E}_{x\sim\rho}\mathbb{E}_{a_{1:T}\sim\pi_{\theta_{\text{old}}}}\left[\sum_{t=1}^{T}\widetilde{A}(s_t,a_t)\min\left\{\psi(s_t,a_t),\text{clip}\left(\psi(s_t,a_t),1-\delta,1+\delta\right)\right\}\right].$$

$$A(s_t,a_t)=\sum_{j=0}^{T-t}\lambda^j\text{advantage}_{t+j}=\sum_{j=0}^{T}\lambda^j[r(s_{t+j},a_{t+j})+\gamma V(s_{t+1+j})-V(s_{t+j})],$$

$\gamma = 1, \lambda = 1$

（最佳实践）

$$\mathcal{L}_{\text{ppo}}(\theta) = \mathbb{E}_{x\sim\rho}\mathbb{E}_{a_{1:T}\sim\pi_\theta}\left[\sum_{t=1}^{T} r(x,a_{1:T}) - V(x,a_{1:t})\right]$$

- 值函数作为基准

REINFORCE变体

- ReMax
- GRPO
- DAPO

- 蒙特卡罗估计基准

[Li, Ziniu, et al. "Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models." *arXiv preprint arXiv:2310.10505* (2023).]

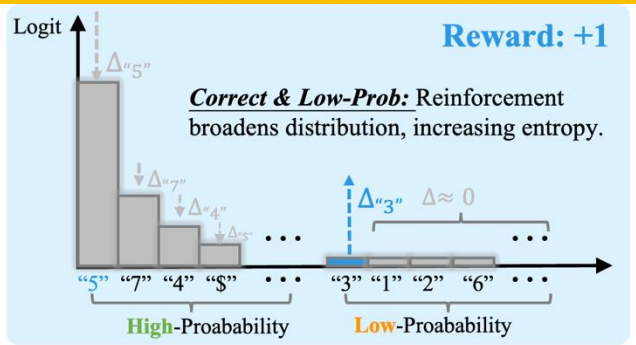
从GRPO到DAPO

强化学习算法常使用近端优化（Proximal Optimization）框架

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

近端优化：限制模型更新区间为 $[\pi(o|q,o_{<t}) * (1-\varepsilon), \pi(o|q,o_{<t}) * (1+\varepsilon)]$

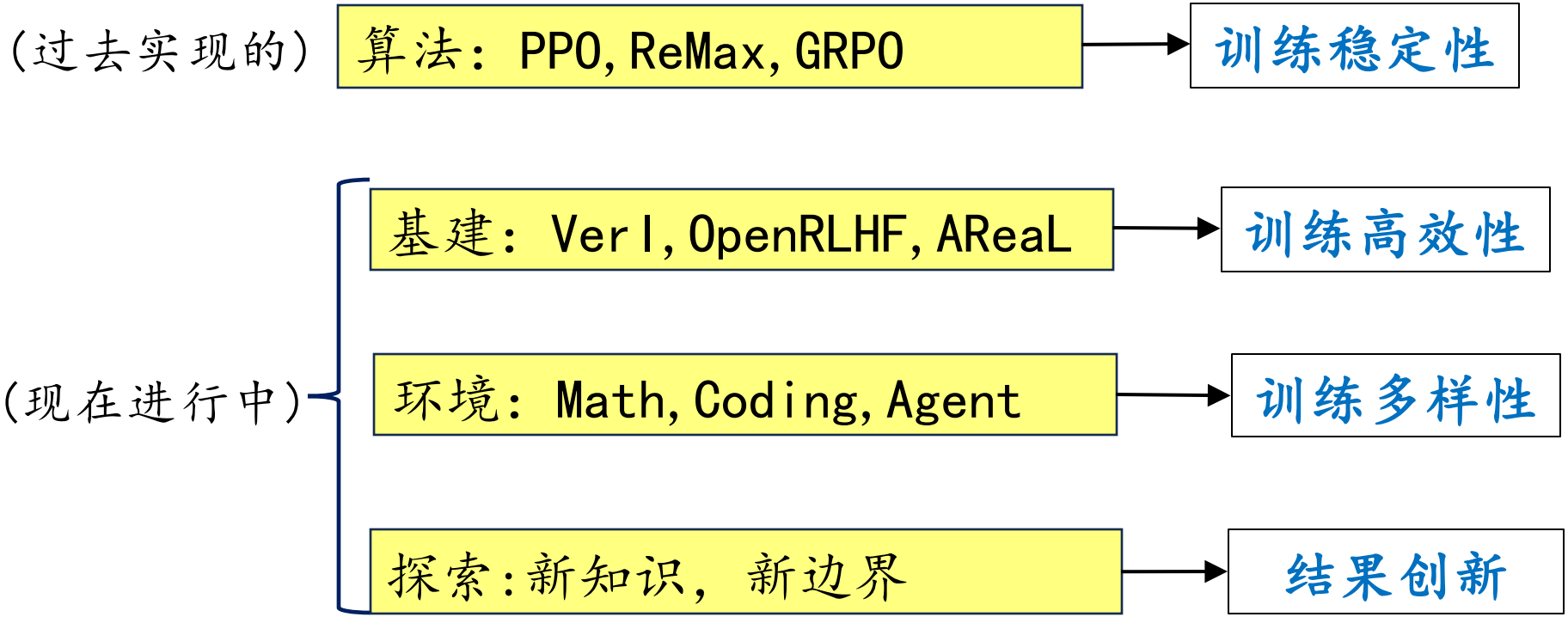
David bought 5 shorts at $7 each, so the total cost was $35.



DAPO：近端优化会额外限制小概率词元的更新，导致稀有轨迹的学习效率较低。因此提出非对称更新，增大右侧的 ε 值

[Yu, Qiying, et al. "Dapo: An open-source llm reinforcement learning system at scale." *arXiv preprint arXiv:2503.14476* (2025).]

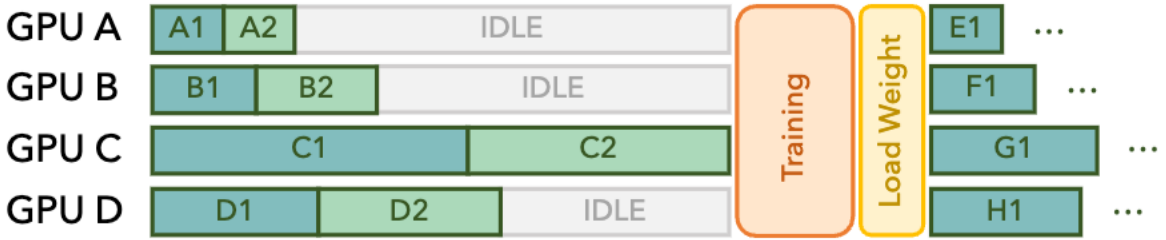
强化学习前沿进展


(过去实现的)
算法：PPO, ReMax, GRPO
训练稳定性
(现在进行中)
基建：Verl, OpenRLHF, AReaL
训练高效性
环境：Math, Coding, Agent
训练多样性
探索：新知识，新边界
结果创新

强化学习框架升级

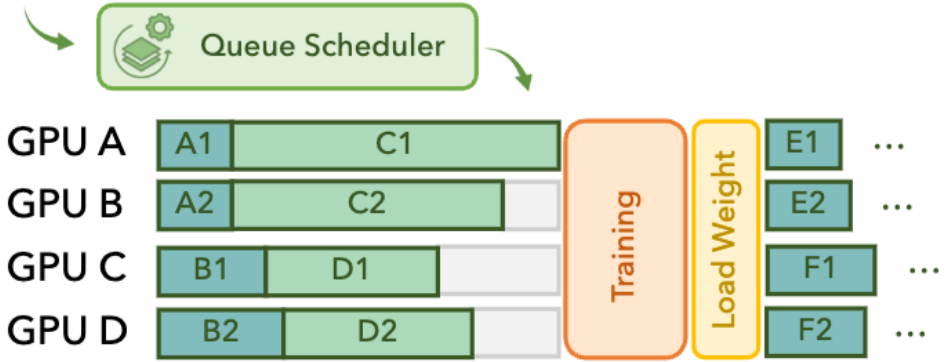
传统“同步”训练框架

GPU利用率低



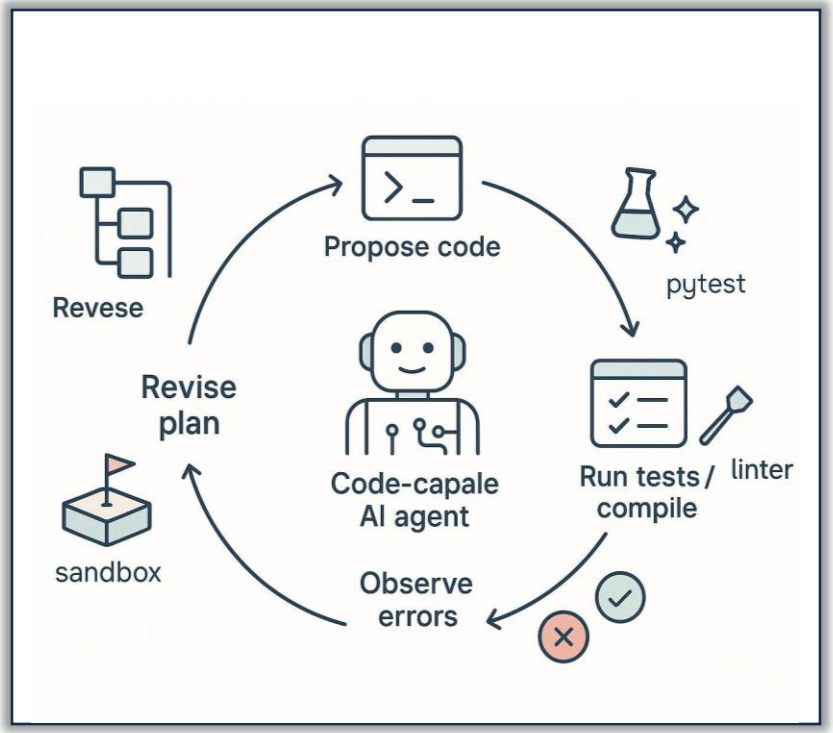
新兴“异步”训练框架

合理调度，GPU资源充分利用

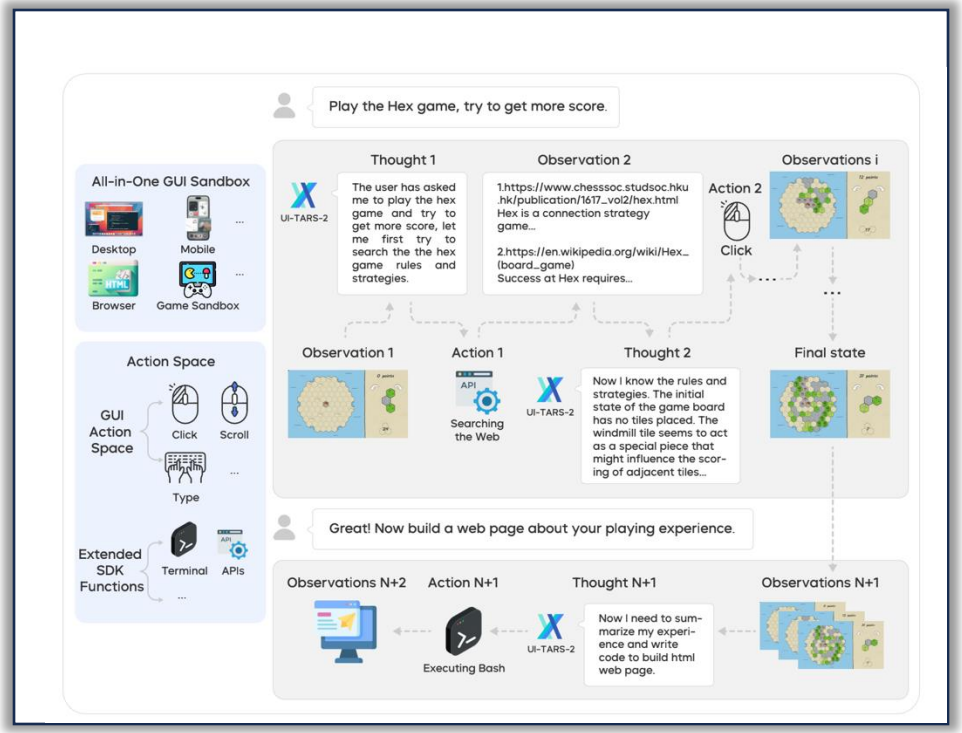


Lu, Han, et al. "Part II: ROLL Flash--Accelerating RLVR and Agentic Training with Asynchrony." *arXiv preprint arXiv:2510.11345* (2025).

强化学习开发生态升级

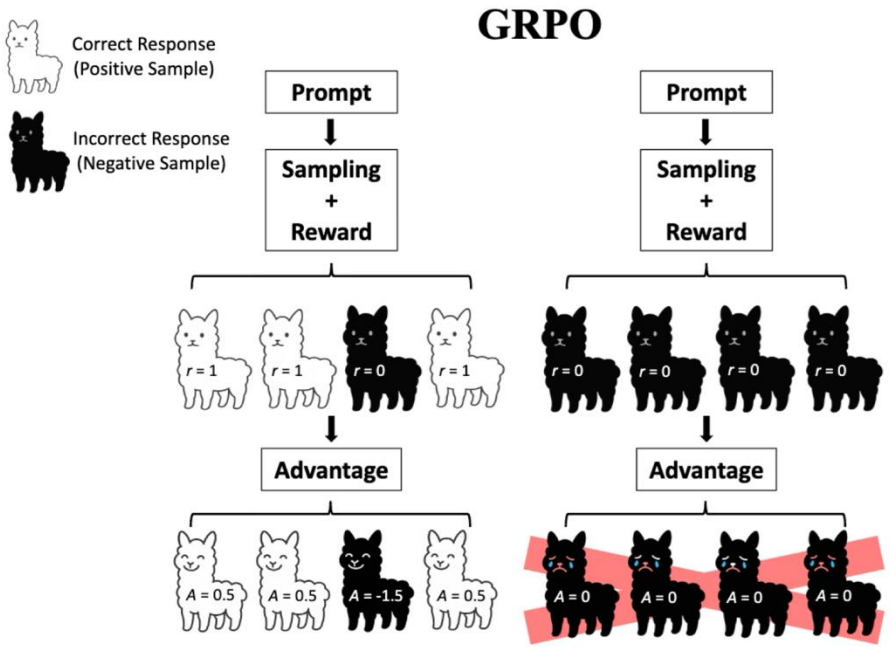


自主编程与测试



多模态交互式训练

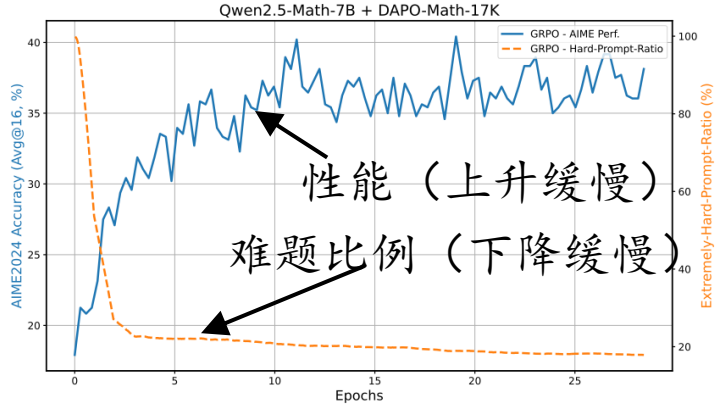
GRPO现有挑战



正负样本混合→梯度非0 → 学习

全负样本→梯度为0 → 丢弃

核心问题：难题上的样本同质，梯度消失



[Chen, Peter, et al. "Spectral Policy Optimization: Coloring your Incorrect Reasoning in GRPO." *arXiv preprint arXiv:2505.11595* (2025).]

[Li, Ziniu, et al. "Knapsack RL: Unlocking Exploration of LLMs via Optimizing Budget Allocation." *arXiv preprint arXiv:2509.25849* (2025).]

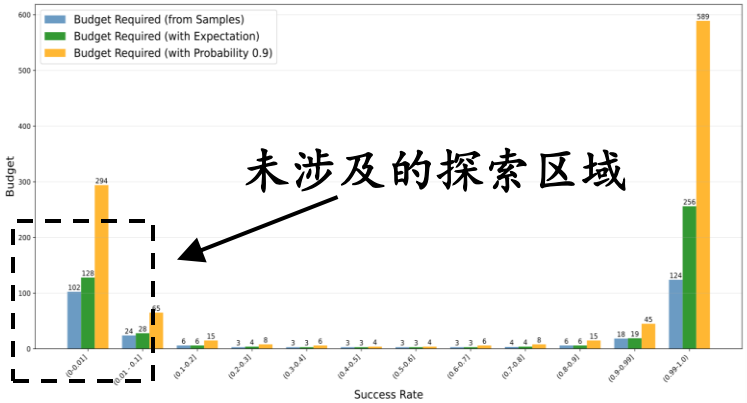
从探索的角度看模型训练

模型收集到的样本都是负的

已有观点 → **模型缺乏知识，真的不会**

新观点 → **模型尝试次数不够多**

经验观察



理论分析

Theorem 1 (Exploration Budget). *Given a prompt i with the success rate $p_i \in (0,1)$, we have that*

- ***High probability bound:*** *For any $\alpha \in (0,1)$, to ensure $\mathbb{P}(g_i \neq 0) \geq \alpha$, it suffices to take $N \gtrsim \frac{\ln(1-\alpha)}{\ln(\max\{p_i, 1-p_i\})}$.*
- ***Expected number of rollouts:*** *Let N_i^{first} denote the number of independent rollouts required until $g_i \neq 0$ is achieved for the first time. Its expectation is: $\mathbb{E}[N_i^{\text{first}}] = 1/p_i + 1/(1-p_i) - 1$.*

[Li, Ziniu, et al. "Knapsack RL: Unlocking Exploration of LLMs via Optimizing Budget Allocation." *arXiv preprint arXiv:2509.25849* (2025).]

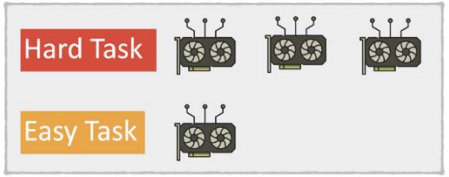
探索资源智能分配

任务无关、均匀的算力分配

- 所有任务获得相同计算预算
- 资源分散，单任务探索深度受限

任务定制化的探索算力分配

- 基于背包算法的智能资源分配
- 难任务获得更多预算，提升整体收益



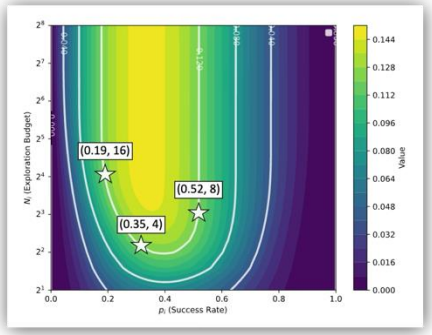
$$\max_{N_1,\ldots,N_M} \sum_{i=1}^{M} \text{Value}(N_i, p_i)$$

$$\text{subject to} \sum_{i=1}^{M} N_i = N_{\text{total}},$$

$$\text{Value}(N_i, p_i) = \text{ProbNonZeroGradient}(N_i, p_i) \times \text{InfoGain}(p_i)$$

$$\text{ProbNonZeroGradient} = 1 - p_i^{N_i} - (1 - p_i)^{N_i}$$

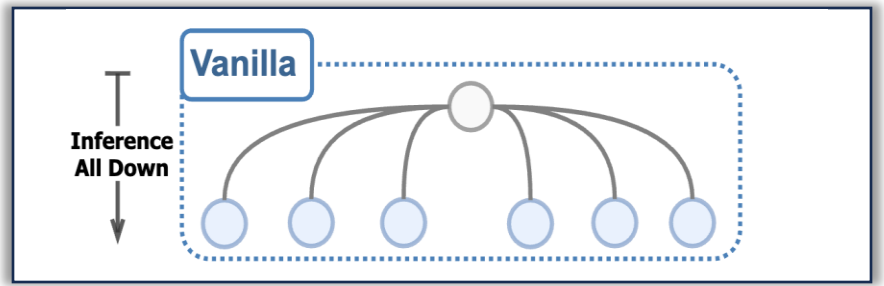
$$\text{InfoGain} = p_i(1 - p_i)^2$$



[Li, Ziniu, et al. "Knapsack RL: Unlocking Exploration of LLMs via Optimizing Budget Allocation." *arXiv preprint arXiv:2509.25849* (2025).]

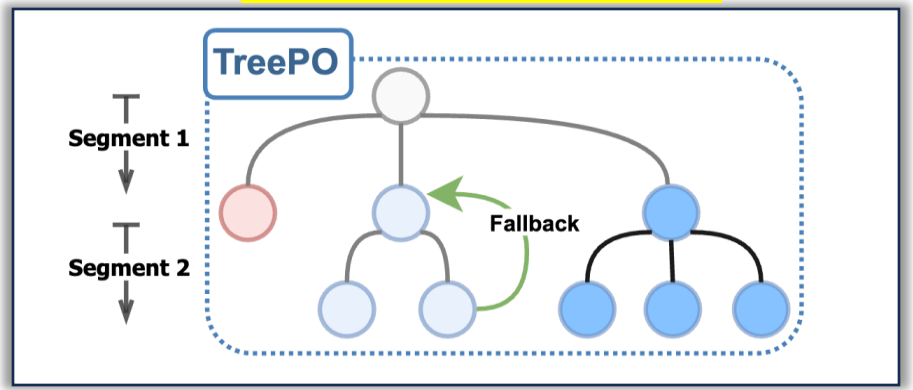
探索方法的进化

传统的在线探索方法



- 每次必须回到“起点”重新开始
- 探索效率低，难以规模化

基于树的结构化探索



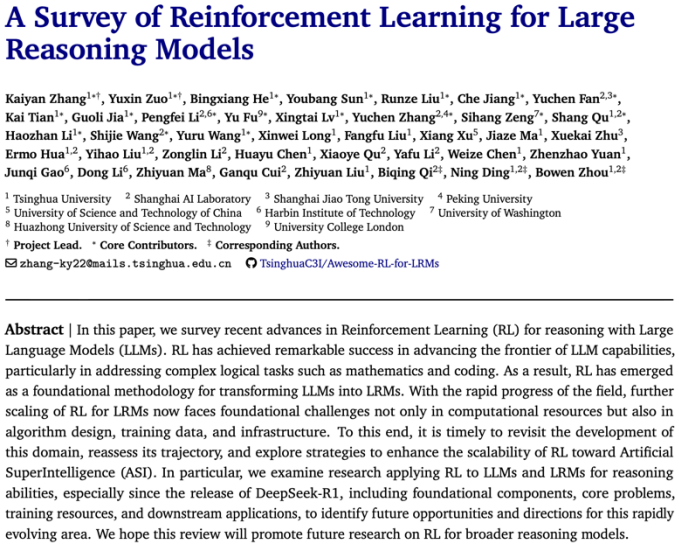
- 利用树结构实现状态回溯与缓存复用
- 相同计算资源下，探索效率提升30%

[Li, Yizhi, et al. "Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling." *arXiv preprint arXiv:2508.17445* (2025).]

总结

- 大模型的发展面临数据量太大、计算量大的挑战，需要Computation Efficient的算法
  - ✓ PPO → RENFORCE
- 大模型面临数据并质，训练不稳定的方向，需要方差小、训练稳定的算法
  - ✓ ReMax → GRPO → DAPO
- 在解决算法和训练的稳定性后，强化学习应更关注探索，挖掘新知识，训练新能力
  - ✓ 智能体训练，异步强化学习
  - ✓ 树搜索，算力背包分配

参考文献



A Survey of Reinforcement Learning for Large Reasoning Models

Kaiyan Zhang[1*†], Yuxin Zuo[1*†], Bingxiang He[1*], Youbang Sun[1*], Runze Liu[1*], Che Jiang[1*], Yuchen Fan[2,3*], Kai Tian[1*], Guoli Jia[1*], Pengfei Li[2,6*], Yu Fu[9*], Xingtai Lv[1*], Yuchen Zhang[2,4*], Sihang Zeng[7*], Shang Qu[1,2*], Haozhan Li[1*], Shijie Wang[2*], Yuru Wang[1*], Xinwei Long[1], Fangfu Liu[1], Xiang Xu[5], Jiaze Ma[1], Xuekai Zhu[3], Ermo Hua[1,2], Yihao Liu[1,2], Zonglin Li[2], Huayu Chen[1], Xiaoye Qu[2], Yafu Li[2], Weize Chen[1], Zhenzhao Yuan[1], Junqi Gao[6], Dong Li[6], Zhiyuan Ma[8], Ganqu Cui[2], Zhiyuan Liu[1], Biqing Qi[2‡], Ning Ding[1,2‡], Bowen Zhou[1,2‡]

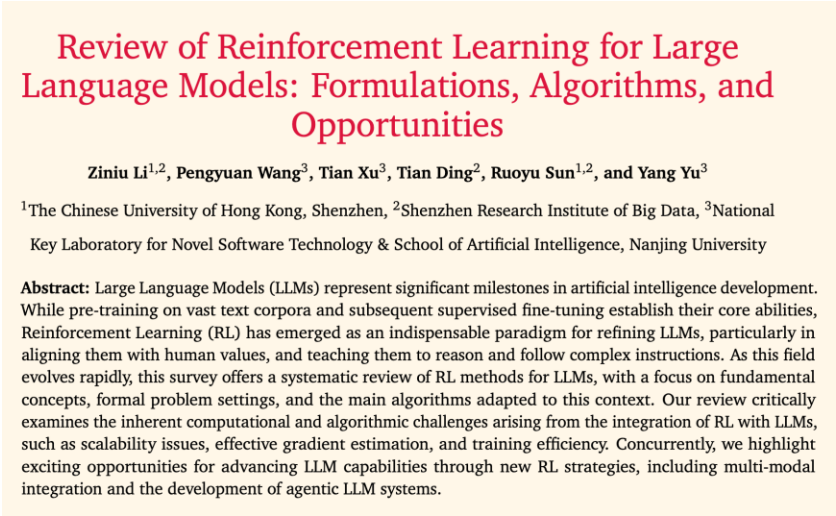
[1] Tsinghua University [2] Shanghai AI Laboratory [3] Shanghai Jiao Tong University [4] Peking University [5] University of Science and Technology of China [6] Harbin Institute of Technology [7] University of Washington [8] Huazhong University of Science and Technology [9] University College London

† Project Lead. * Core Contributors. ‡ Corresponding Authors.

zhang-ky22@mails.tsinghua.edu.cn TsinghuaC3I/Awesome-RL-for-LRMs

Abstract | In this paper, we survey recent advances in Reinforcement Learning (RL) for reasoning with Large Language Models (LLMs). RL has achieved remarkable success in advancing the frontier of LLM capabilities, particularly in addressing complex logical tasks such as mathematics and coding. As a result, RL has emerged as a foundational methodology for transforming LLMs into LRMs. With the rapid progress of the field, further scaling of RL for LRMs now faces foundational challenges not only in computational resources but also in algorithm design, training data, and infrastructure. To this end, it is timely to revisit the development of this domain, reassess its trajectory, and explore strategies to enhance the scalability of RL toward Artificial SuperIntelligence (ASI). In particular, we examine research applying RL to LLMs and LRMs for reasoning abilities, especially since the release of DeepSeek-R1, including foundational components, core problems, training resources, and downstream applications, to identify future opportunities and directions for this rapidly evolving area. We hope this review will promote future research on RL for broader reasoning models.

[https://arxiv.org/abs/2509.08827]



Review of Reinforcement Learning for Large Language Models: Formulations, Algorithms, and Opportunities

Ziniu Li[1,2], Pengyuan Wang[3], Tian Xu[3], Tian Ding[2], Ruoyu Sun[1,2], and Yang Yu[3]

[1]The Chinese University of Hong Kong, Shenzhen, [2]Shenzhen Research Institute of Big Data, [3]National Key Laboratory for Novel Software Technology & School of Artificial Intelligence, Nanjing University

Abstract: Large Language Models (LLMs) represent significant milestones in artificial intelligence development. While pre-training on vast text corpora and subsequent supervised fine-tuning establish their core abilities, Reinforcement Learning (RL) has emerged as an indispensable paradigm for refining LLMs, particularly in aligning them with human values, and teaching them to reason and follow complex instructions. As this field evolves rapidly, this survey offers a systematic review of RL methods for LLMs, with a focus on fundamental concepts, formal problem settings, and the main algorithms adapted to this context. Our review critically examines the inherent computational and algorithmic challenges arising from the integration of RL with LLMs, such as scalability issues, effective gradient estimation, and training efficiency. Concurrently, we highlight exciting opportunities for advancing LLM capabilities through new RL strategies, including multi-modal integration and the development of agentic LLM systems.

[https://openreview.net/forum?id=ghQQNjSxJc]

- Wang, Peng-Yuan, et al. "A Survey on Large Language Models for Mathematical Reasoning." *arXiv preprint arXiv:2506.08446* (2025).

参考文献

- Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).
- Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022): 27730-27744.
- Li, Ziniu, et al. "Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models." arXiv preprint arXiv:2310.10505 (2023).
- Shao, Zhihong, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." arXiv preprint arXiv:2402.03300 (2024).
- Yu, Qiying, et al. "Dapo: An open-source llm reinforcement learning system at scale." arXiv preprint arXiv:2503.14476 (2025).
- Li, Ziniu, et al. "Knapsack RL: Unlocking Exploration of LLMs via Optimizing Budget Allocation." *arXiv preprint arXiv:2509.25849* (2025).

谢谢