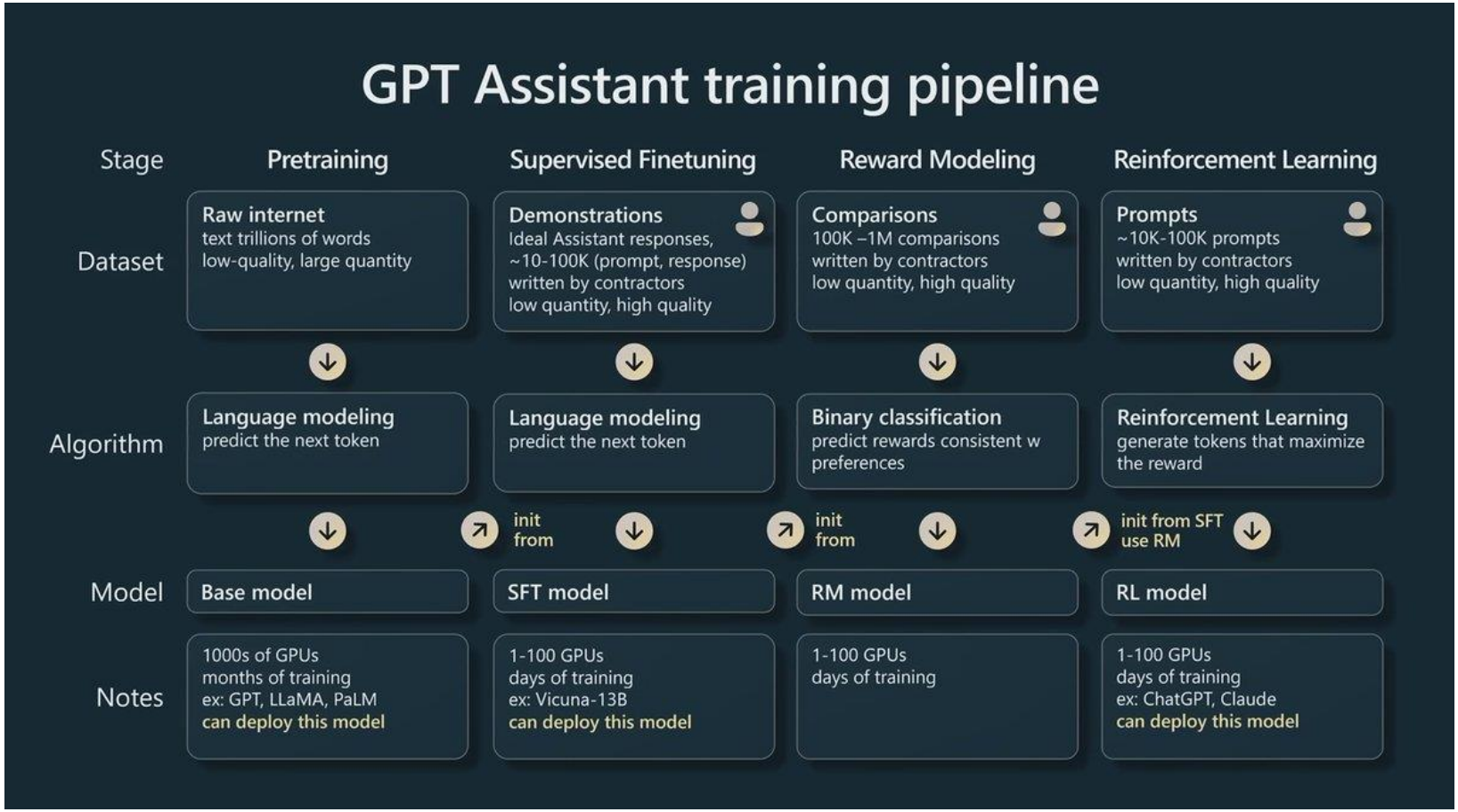
LAMDA
Learning And Mining from DatA
http://www.lamda.nju.edu.cn

南京大学

强化学习驱动的LLM优化：最新进展分享

庞竟成
南京大学

LLM经典训练流程



Pretraining: 嵌入知识 ⇒ SFT: 解锁知识 ⇒ RLHF: 偏好对齐

Figure is from https://weaviate.io/papers/paper20

LLM优化为什么需要RL?

ChatGPT 为什么不用 Reward-Model 的数据直接 fine-tune，而用 RL...

俞扬
2023 年度新知答主

已关注

Ke Xue 等 522 人赞同了该回答

2023-04-19：

这是一个很好的问题

另一个帖子里子牛同学已经说到在RL领域里与reward学习等价的apprentice learning 要比与fine-tune等价的bc好不少。（基于ranking loss学的reward其实要比irl学的差，但是工程上更容易）

基于reward学习的RL训练可比监督学习fine tune好不少，这一现象我们在RL的研究中早已观察到，特别是在OOD数据上有显著差别。最近的分析取得了一些进展，待公开后更新。

LLM优化为什么需要RL?

UNDERSTANDING THE EFFECTS OF RLHF ON LLM GENERALISATION AND DIVERSITY

Robert Kirk* [α] **Ishita Mediratta** [β] **Christoforos Nalmpantis** [β] **Jelena Luketina** [γ]
Eric Hambro [β] **Edward Grefenstette** [α] **Roberta Raileanu** [β]

[α] University College London, [β] Meta, [γ] University of Oxford

ABSTRACT

Large language models (LLMs) fine-tuned with reinforcement learning from human feedback (RLHF) have been used in some of the most widely deployed AI models to date, such as OpenAI's ChatGPT or Anthropic's Claude. While there has been significant work developing these methods, our understanding of the benefits and downsides of each stage in RLHF is still limited. To fill this gap, we present an extensive analysis of how each stage of the process (i.e. supervised fine-tuning (SFT), reward modelling, and RLHF) affects two key properties: out-of-distribution (OOD) generalisation and output diversity. OOD generalisation is crucial given the wide range of real-world scenarios in which these models are being used, while output diversity refers to the model's ability to generate varied outputs and is important for a variety of use cases. We perform our analysis across two base models on both summarisation and instruction following tasks, the latter being highly relevant for current LLM use cases. We find that RLHF generalises better than SFT to new inputs, particularly as the distribution shift between train and test becomes larger. However, RLHF significantly reduces output diversity compared to SFT across a variety of measures, implying a tradeoff in current LLM fine-tuning methods between generalisation and diversity. Our results provide guidance on which fine-tuning method should be used depending on the application, and show that more research is needed to improve the tradeoff between generalisation and diversity.

models on both summarisation and instruction following tasks, the latter being highly relevant for current LLM use cases. We find that RLHF generalises better than SFT to new inputs, particularly as the distribution shift between train and test becomes larger. However, RLHF significantly reduces output diversity compared to

经RL优化后的LLM展现出更强的泛化性

这意味着大多数新模型的行为都是 RLHF 的产物。

那么让我们看看 RLHF 触发的能力：

- **翔实的回应**：text-davinci-003 的生成通常比 text-davinci-002 长。ChatGPT 的回应则更加冗长，以至于用户必须明确要求“用一句话回答我”，才能得到更加简洁的回答。这是 RLHF 的直接产物。
- **公正的回应**：ChatGPT 通常对涉及多个实体利益的事件（例如政治事件）给出非常平衡的回答。这也是RLHF的产物。
- **拒绝不当问题**：这是内容过滤器和由 RLHF 触发的模型自身能力的结合，过滤器过滤掉一部分，然后模型再拒绝一部分。
- **拒绝其知识范围之外的问题**：例如，拒绝在2021 年 6 月之后发生的新事件（因为它没在这之后的数据上训练过）。这是 RLHF 最神奇的部分，因为它使模型能够隐式地区分哪些问题在其知识范围内，哪些问题不在其知识范围内。

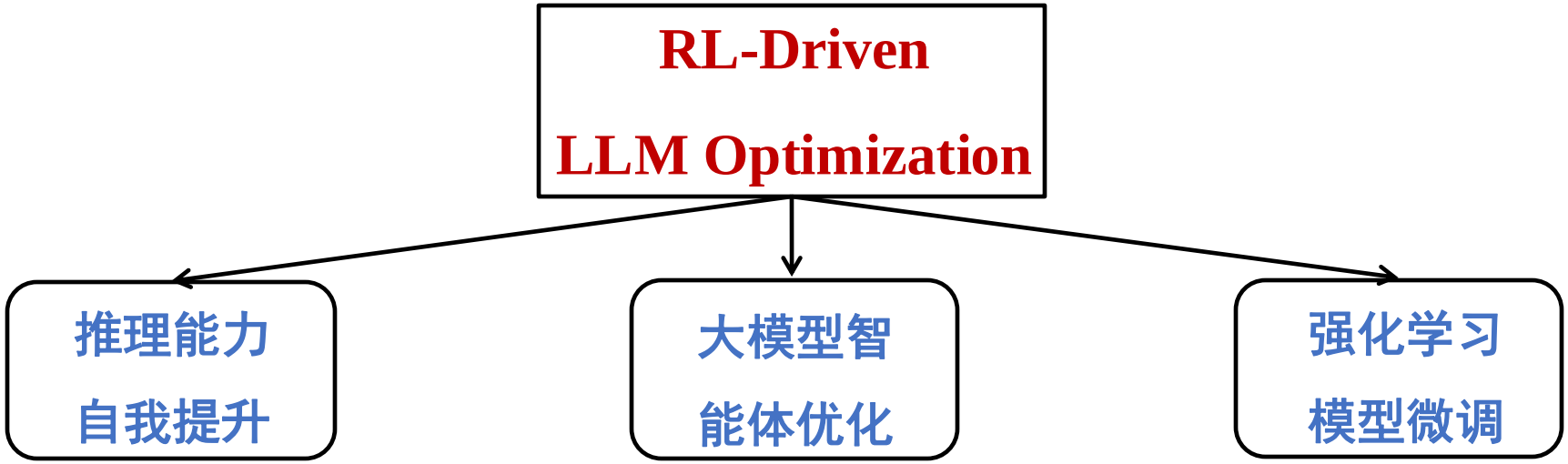
学会“拒绝回答”的能力
(优化一些难提供示例的特性)

Left: Understanding the Effects of RLHF on LLM Generalisation and Diversity. Kirk, et al. ICLR 2024.
Right: Drawn from https://yaofu.notion.site/GPT-3-5-360081d91ec245f29029d37b54573756.

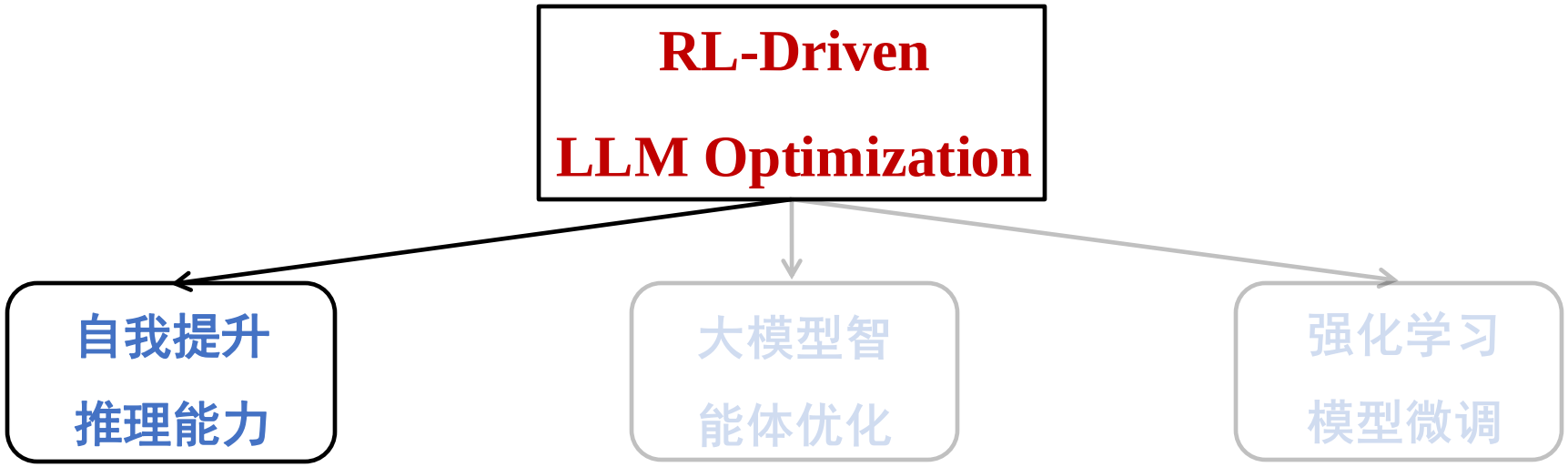
RL驱动的LLM优化前沿问题

RL用于LLM优化的场景

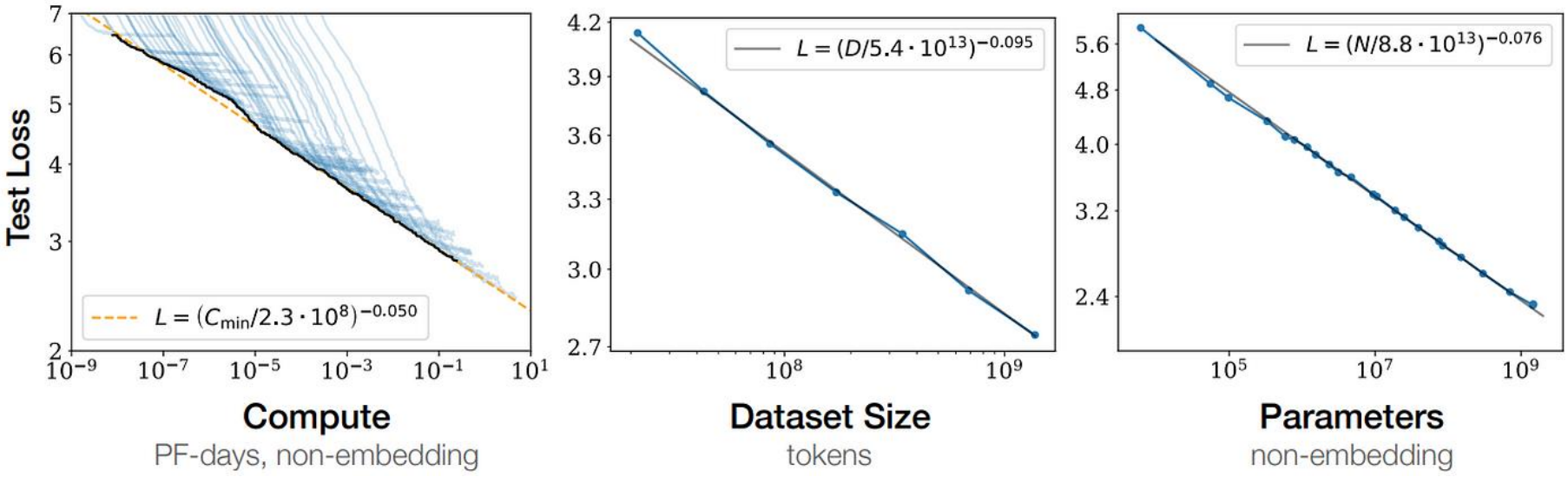
1. **模型自我提升：** 突破训练数据，实现模型基座自我提升。
2. **大模型智能体：** 大模型智能体如何从环境反馈中学习？
3. **强化学习微调：** 尝试多路径思考，从标注获取奖励信号。



RL驱动的LLM优化前沿问题



大模型自我提升

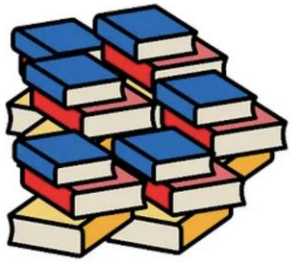


当前大模型能力提升主要源于数据的规模化（Scaling）

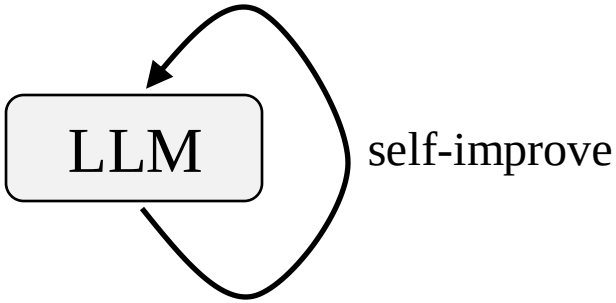
大模型自我提升

如何使大模型能力进一步提升？

更多数据

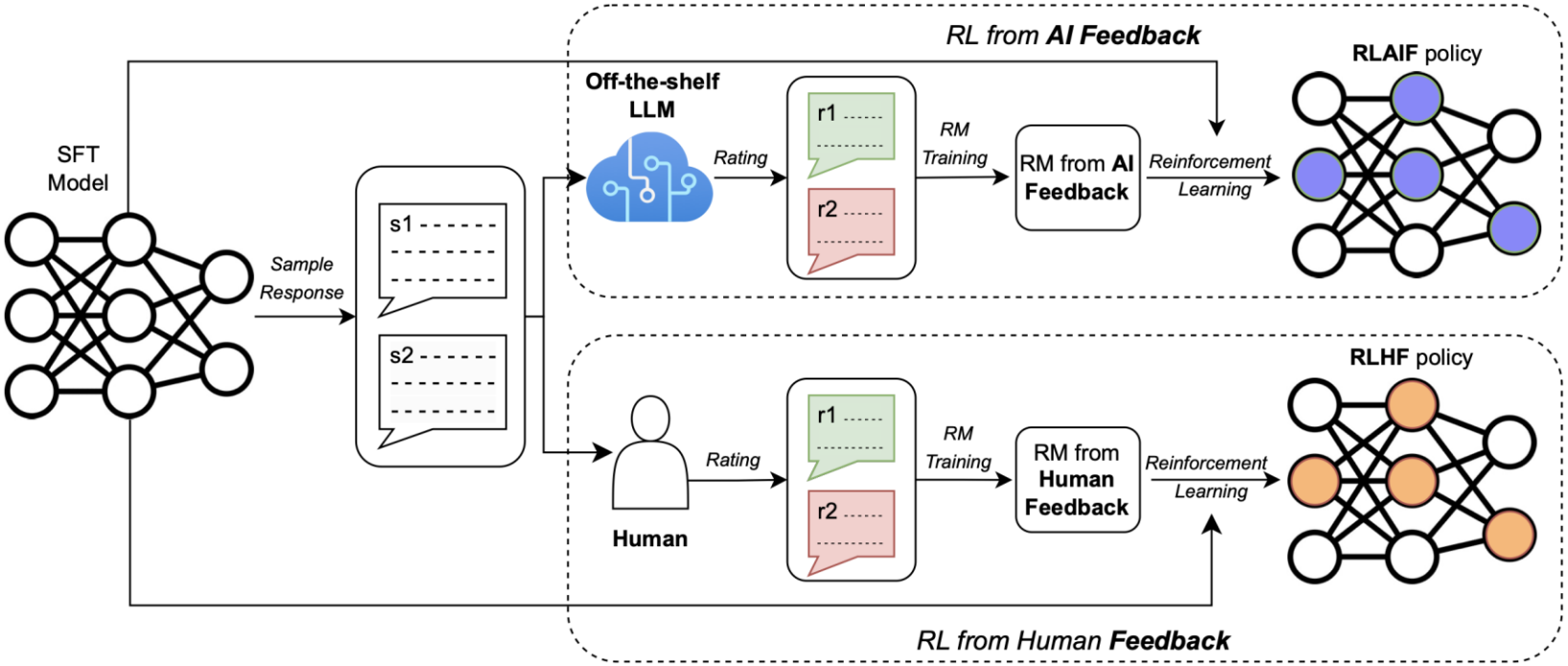


自我提升



大模型自我提升

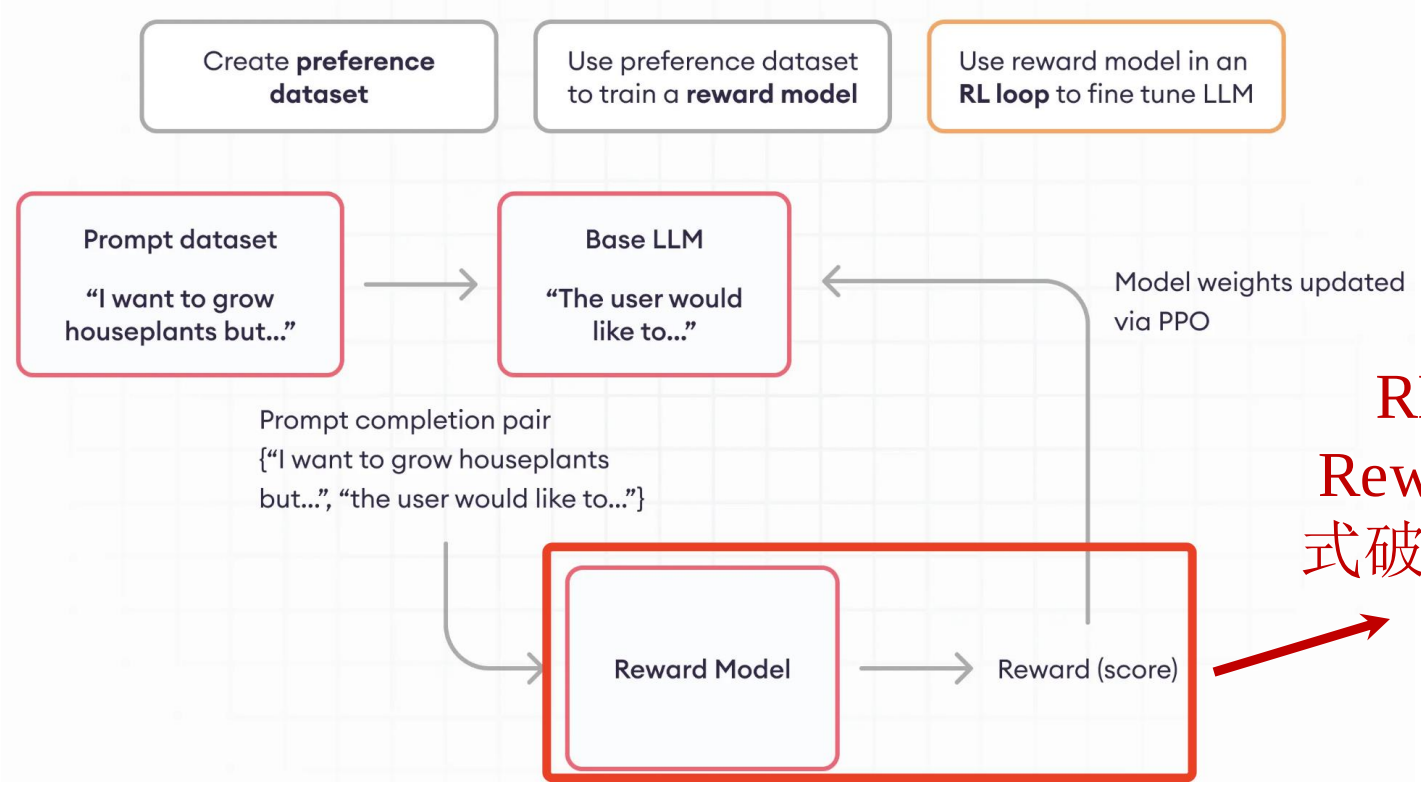
经典方法RLAIF: AI代替RLHF中的人类工作



问题： 大语言模型自我提升原理不清晰；
额外训练判别式奖赏模型，工作流程繁琐。

RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. Lee, et al. ICML 2024.

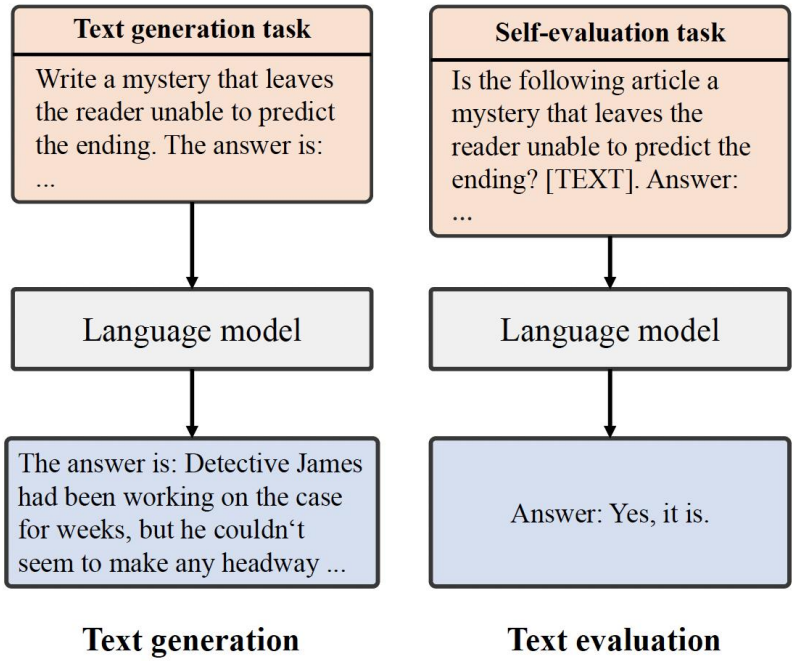
大模型自我提升

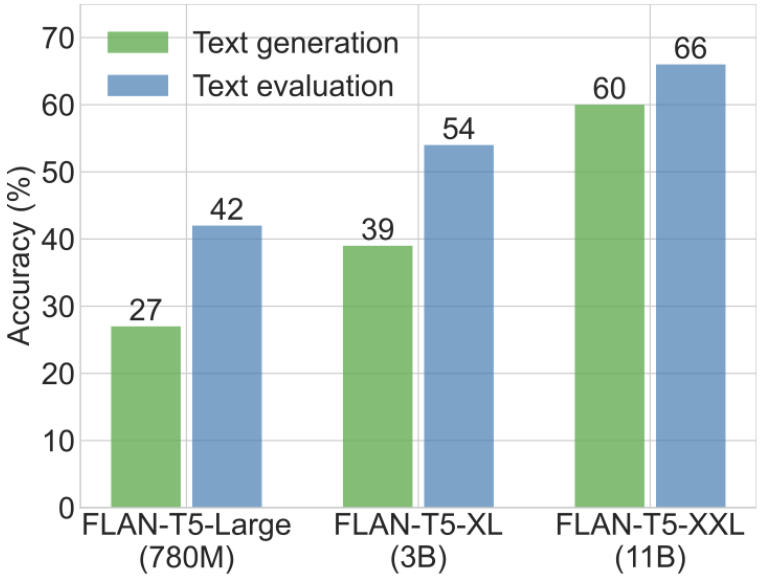


RLHF/RLAIF的Reward Modelling方式破坏了预训练LLM泛化性

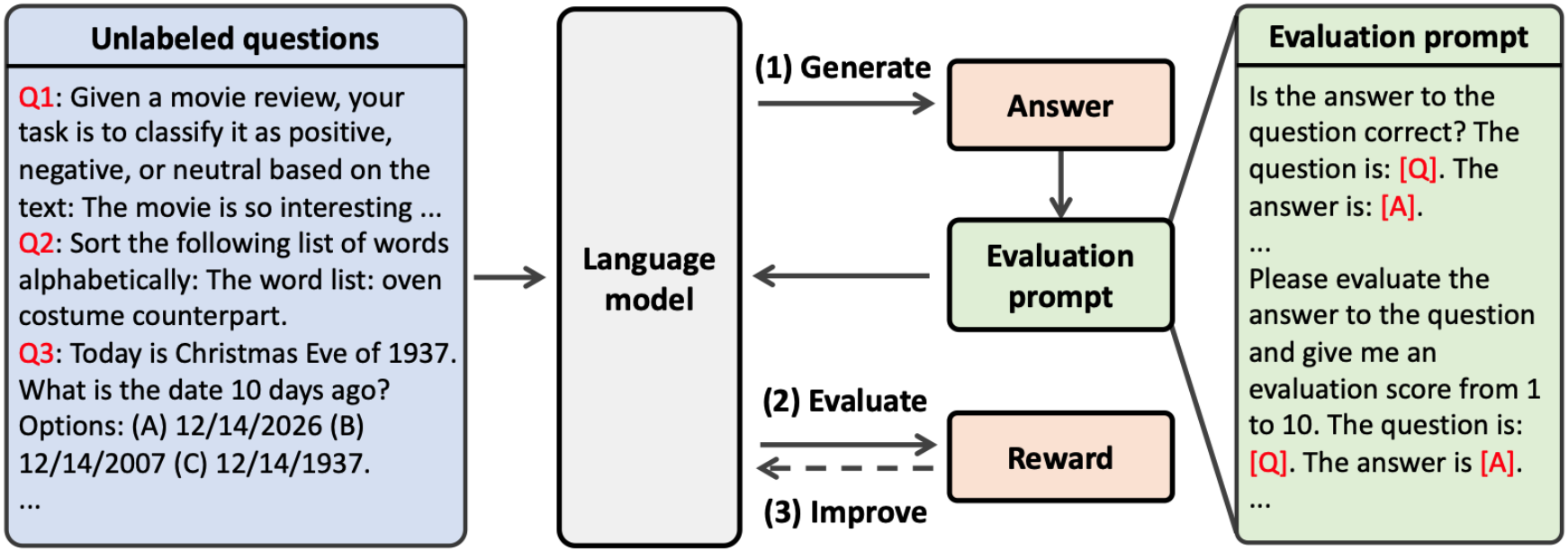
RLC：结合RL的自我提升

ICLR'24：文本评估显著易于文本生成





RLC：结合RL的自我提升



RLC：基于**生成式奖赏**和**强化学习**的大模型自我提升方案

RLC：结合RL的自我提升

Google DeepMind

2024-10-14

Generative Verifiers: Reward Modeling as Next-Token Prediction

Lunjun Zhang[1,2], Arian Hosseini[1,3*], Hritik Bansal[1,4*], Mehran Kazemi[1], Aviral Kumar[1,5] and Rishabh Agarwal[1]
[*]Core Contribution, [1]Google DeepMind, [2]University of Toronto, [3]Mila, [4]UCLA, [5]Carnegie Mellon University

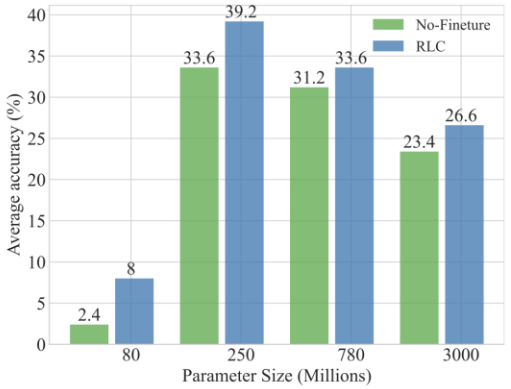
Verifiers or reward models are often used to enhance the reasoning performance of large language models (LLMs). A common approach is the Best-of-N method, where N candidate solutions generated by the LLM are ranked by a verifier, and the best one is selected. While LLM-based verifiers are typically trained as discriminative classifiers to score solutions, they do *not* utilize the text generation capabilities of pretrained LLMs. To overcome this limitation, we instead propose training verifiers using the ubiquitous next-token prediction objective, jointly on verification and solution generation. Compared to standard verifiers, such generative verifiers (GenRM) can benefit from several advantages of LLMs: they integrate seamlessly with instruction tuning, enable chain-of-thought reasoning, and can utilize additional test-time compute via majority voting for better verification. We demonstrate that GenRM outperforms discriminative, DPO verifiers, and LLM-as-a-Judge, resulting in a 16 – 40% improvement in the number of problems solved with Best-of-N on algorithmic and math reasoning tasks. Furthermore, we find that training GenRM with synthetic verification rationales is sufficient to pick out subtle errors on math problems. Finally, we demonstrate that generative verifiers scale favorably with model size and inference-time compute.

Google在15个月后提出“**生成式奖励模型**”LLM训练方案

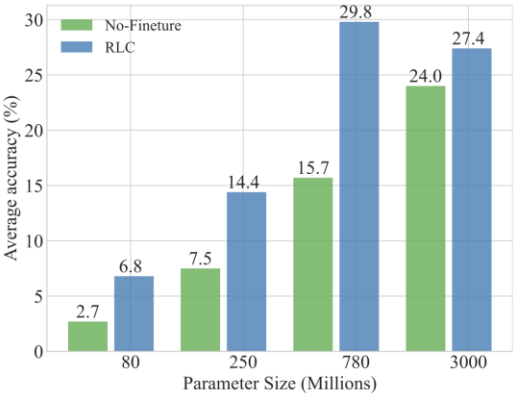
RLC：结合RL的自我提升

	Reasoning about Colored Objects	Logical Deduction (7)	Tracking Shuffled Objects (5)	Object Counting	Tracking Shuffled Objects (3)	Geometric Shapes
RLFT	32.1%	45.7%	12.4%	42.6%	33.6%	18.9%
DG	32.0%	35.2%	12.4%	31.9%	31.2%	5.2%
SC	**39.6%**	27.6%	12.4%	24.0%	33.6%	15.6%
Self-train	19.5%	13.1%	**15.5%**	11.7%	33.1%	12.4%
Self-refine	25.2%	13.2%	8.0%	18.0%	25.2%	10.0%
Best-of-N	26.8%	12.8%	12.1%	14.0%	30.0%	8.4%
RLAIF	30.4%	36.9%	11.4%	32.5%	32.8%	14.0%
RLC	35.0%	**39.2%**	12.2%	**35.4%**	**33.6%**	**17.8%**

	Web of Lies	Sports Understanding	Logical Deduction (3)	Logical Deduction (5)	Penguins in a Table	Navigate
RLFT	72.2%	68.8%	58.6%	41.9%	44.2%	55.6%
DG	43.6%	53.2%	39.6%	28.4%	15.7%	46.4%
SC	48.4%	53.6%	42.8%	30.8%	**35.2%**	**62.8%**
Self-train	51.1%	51.1%	34.0%	18.4%	19.7%	48.7%
Self-refine	47.2%	50.0%	28.4%	17.2%	17.8%	46.0%
Best-of-N	50.0%	**59.2%**	42.0%	22.0%	17.8%	45.2%
RLAIF	52.1%	56.1%	22.0%	33.7%	19.8%	48.8%
RLC	**52.9%**	53.5%	**44.0%**	**34.6%**	29.8%	57.1%



(a) Tracking Shuffled Objects (3)



(b) Penguins in a Table

方案超过Google RLAIF方法，且在小模型上验证效果

自我提升来自哪里?

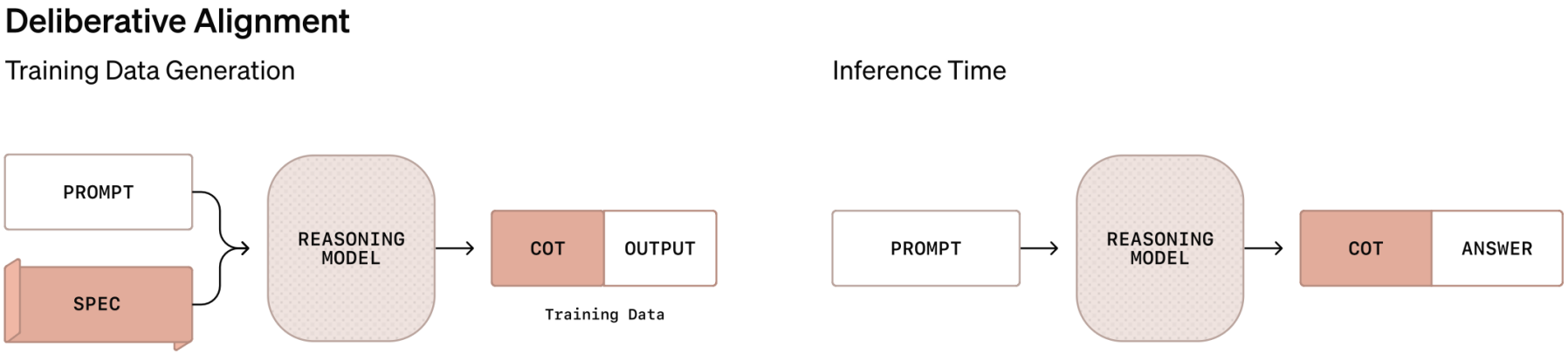
解锁了LLM中的hidden knowledge

We refer to **sharpening** as any process that tilts π_{base} toward responses that are more certain in the sense that they enjoy greater self-reward r_{self}. More formally, a sharpened model $\widehat{\pi}$ is one that (approximately) maximizes the self-reward:

$$\widehat{\pi}(x) \approx \arg\max_{y \in \mathcal{Y}} r_{\text{self}}(y \mid x; \pi_{\text{base}}) \tag{1}$$

直接decode出的答案不一定是最优回答

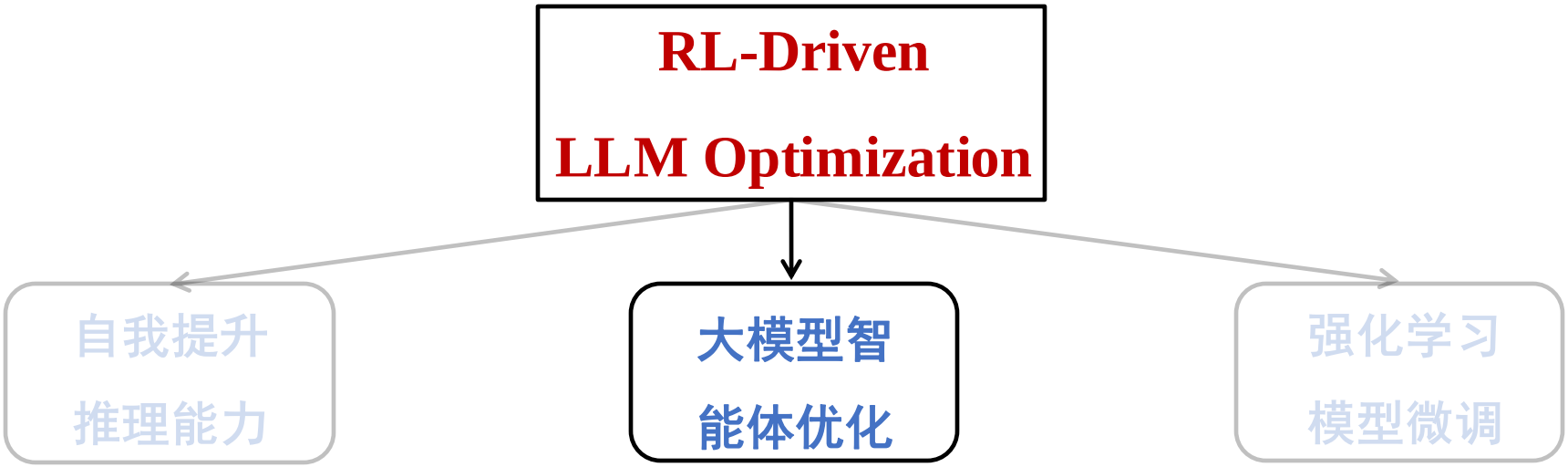
自我提升的下一步



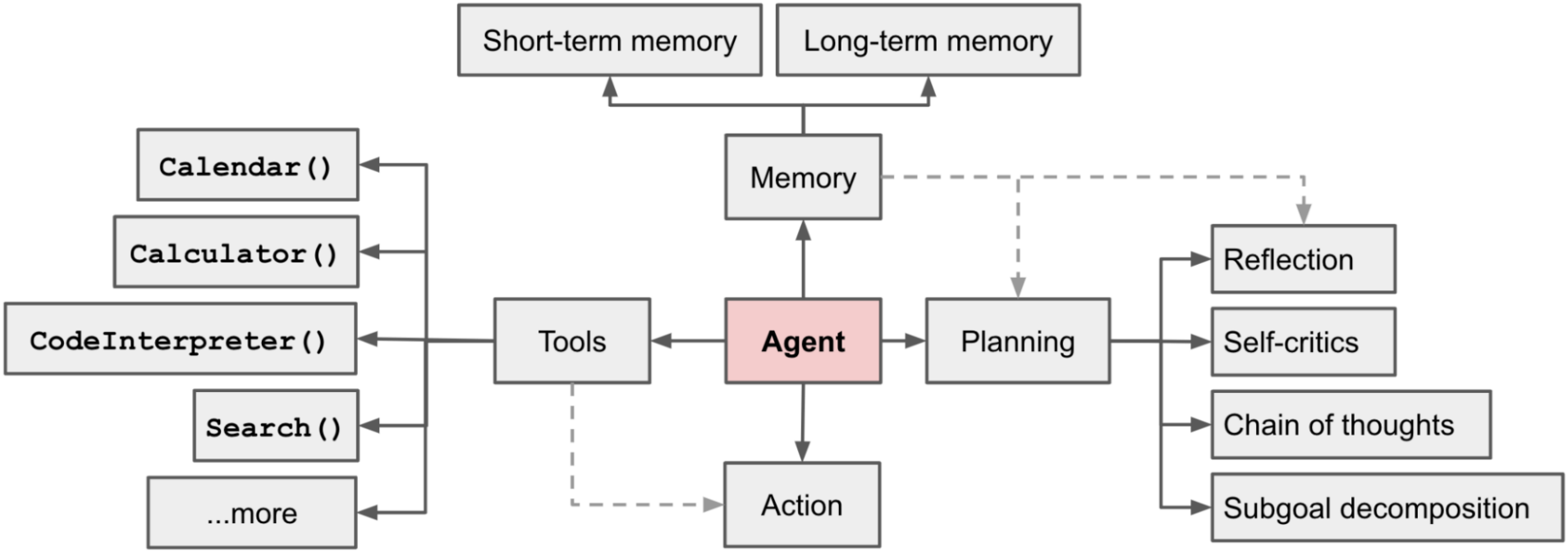
通过LLM长时间思考产生高质量数据作为标注，结合**强化学习微调**进一步优化模型

Figure is from https://openai.com/index/deliberative-alignment/

RL驱动的LLM优化前沿问题



大模型智能体

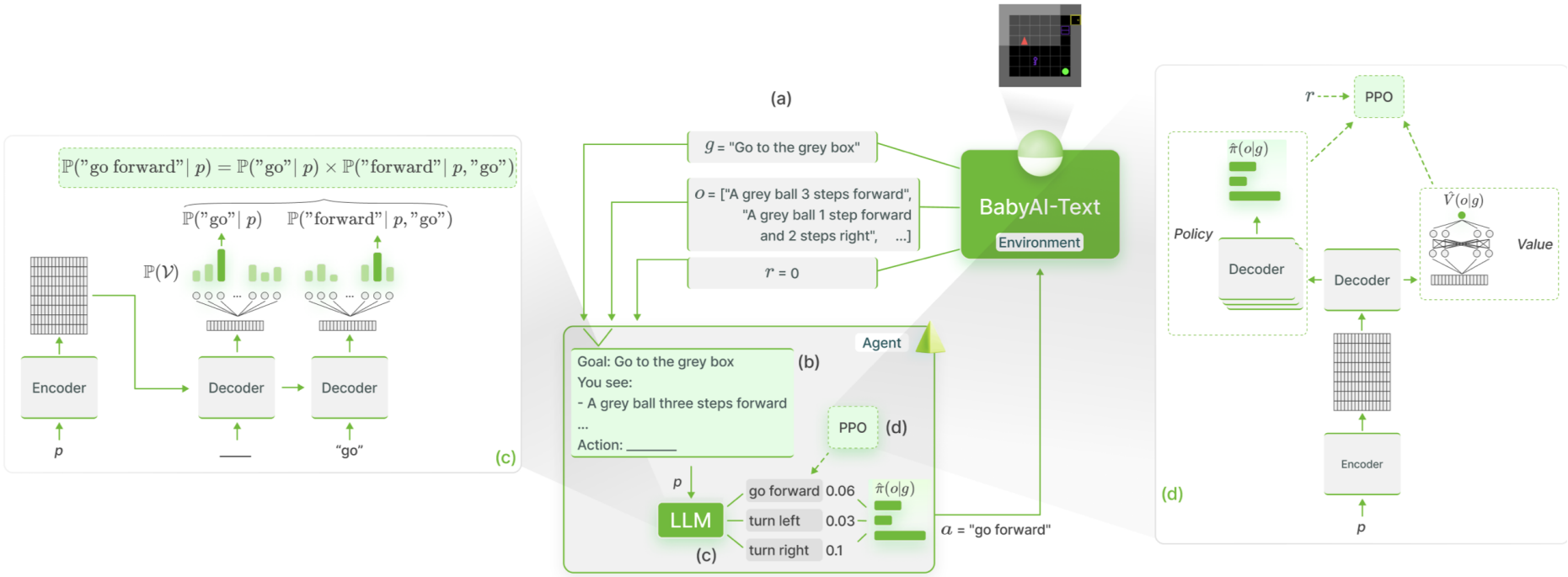


LLM的通用性、推理能力使其天然适合担当“大脑”

关键问题：如何从反馈中学习

LLM Powered Autonomous Agents. Lilian Weng.

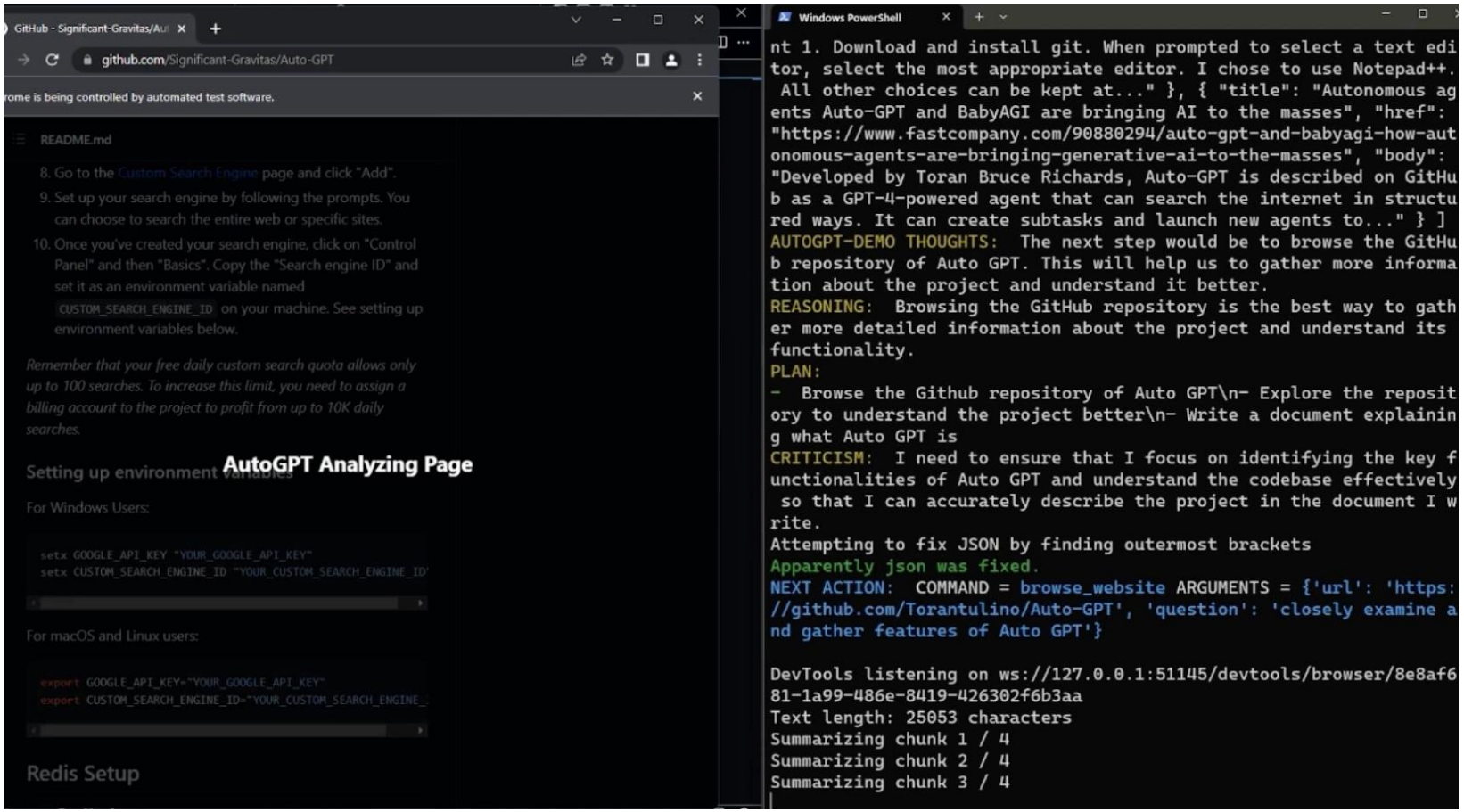
大模型智能体



GLAM使用LLM完成文本游戏BabyAI、

强化学习直接优化模型

大模型智能体

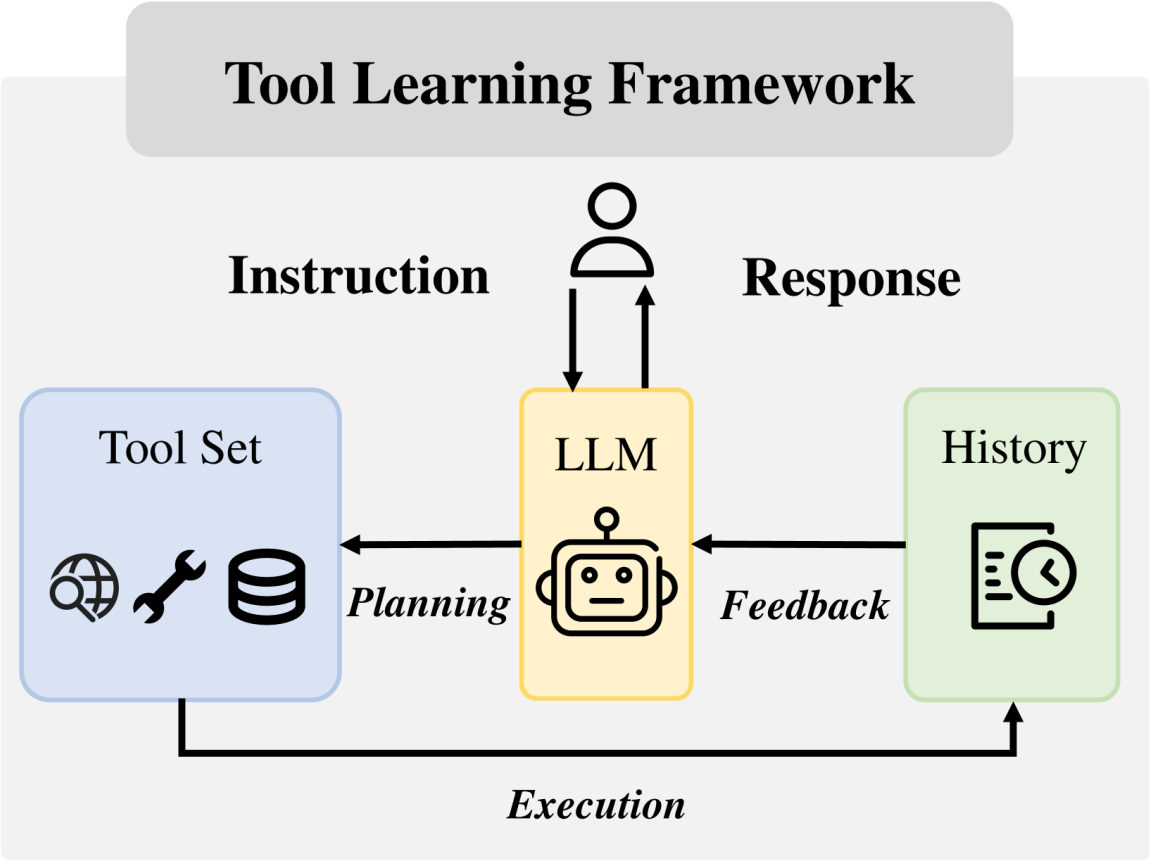


AutoGPT自动操控命令行分析网页

存储海量文本经验

AutoGPT. https://github.com/Significant-Gravitas/AutoGPT

大模型智能体

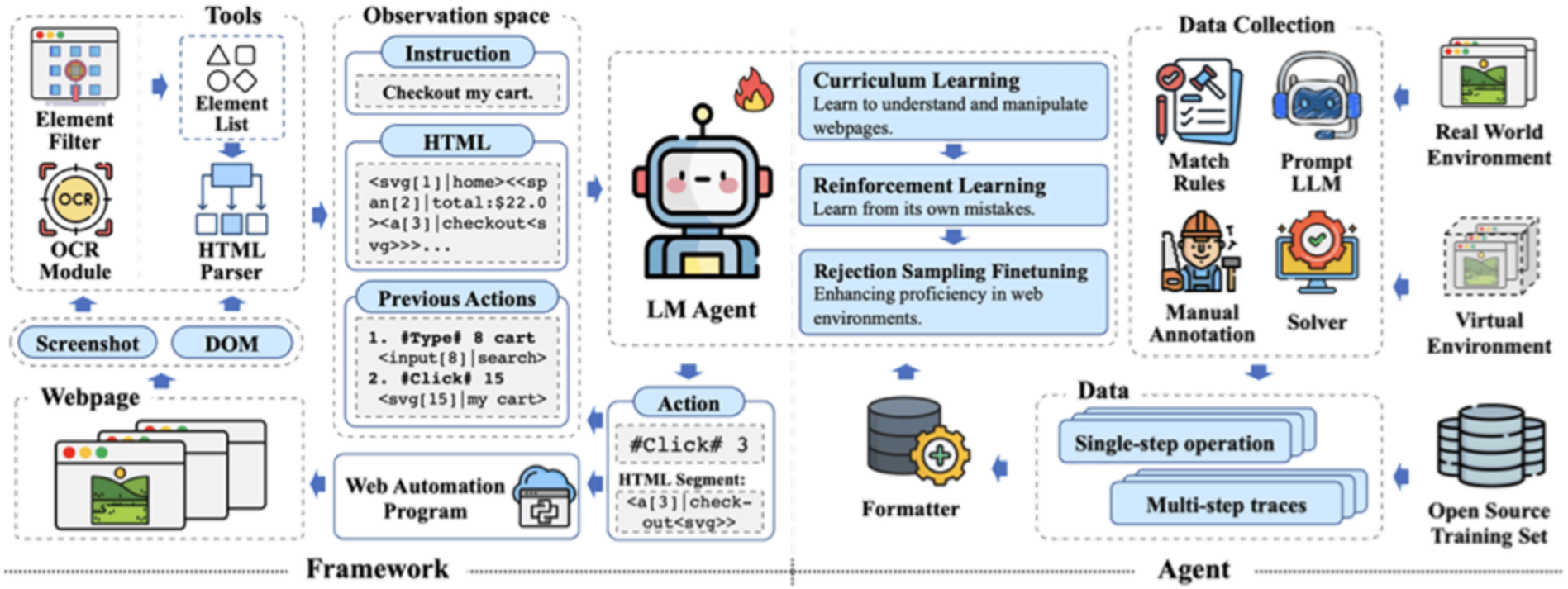


API调用同时操控多种工具完成不同任务

存储文本经验提升任务理解

From Summary to Action: Enhancing Large Language Models for Complex Tasks with Open World APIs. Liu et al.

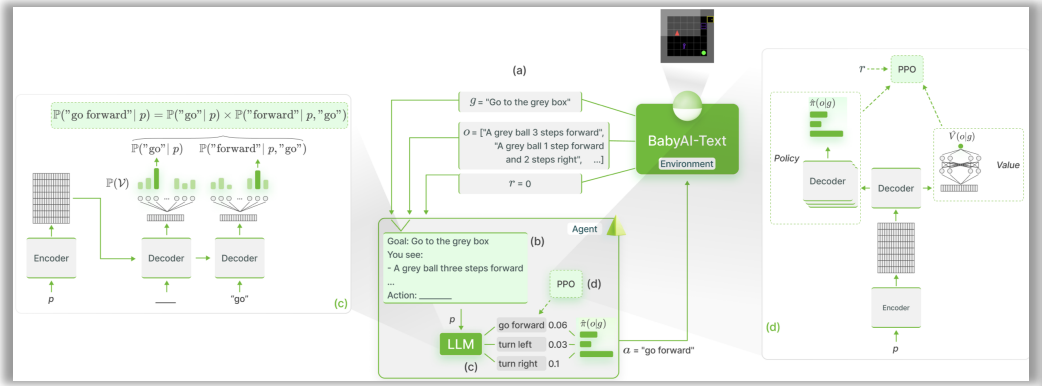
大模型智能体

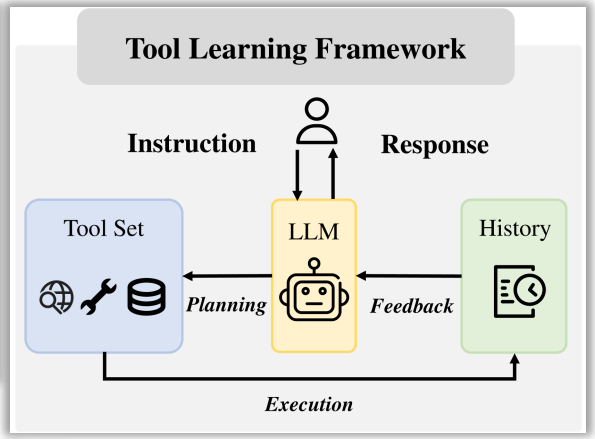


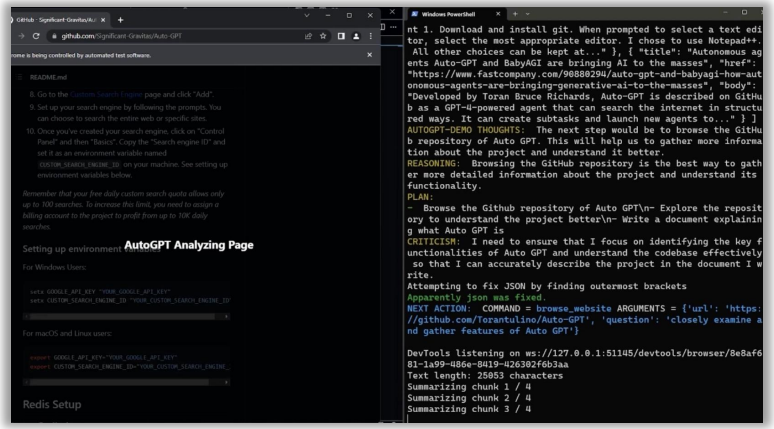
AutoGLM-Phone在移动设备上整合各种工具调用，
使用RL基于环境反馈优化LLM

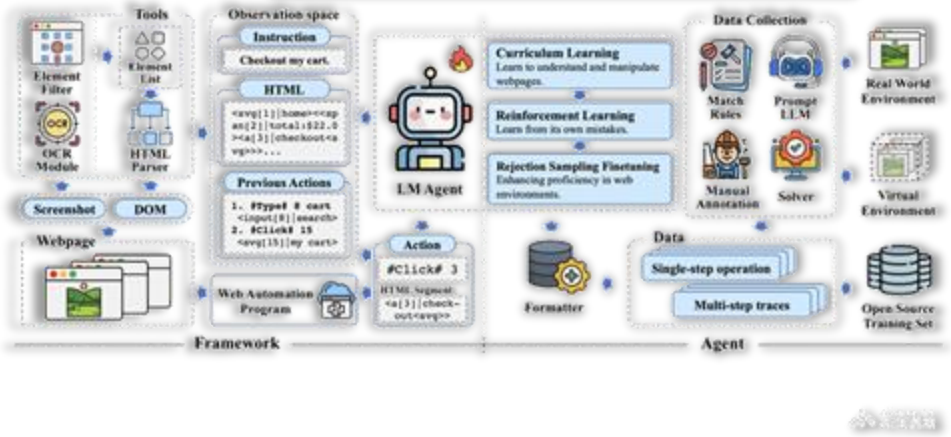
Image is from https://ai-bot.cn/autoglm-web/.

大模型智能体



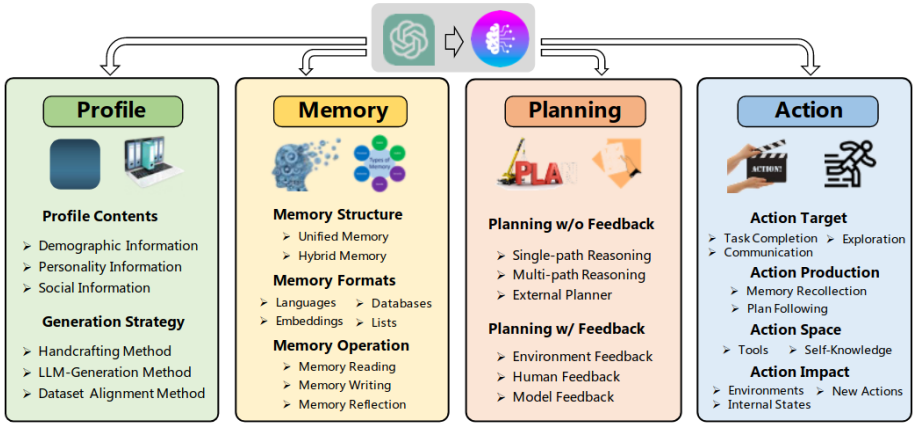






环境反馈难以有效用于提升LLM决策能力

LLM 与 RL构建决策智能体



LLM agent[1]

RL agent

通用知识
低成功率

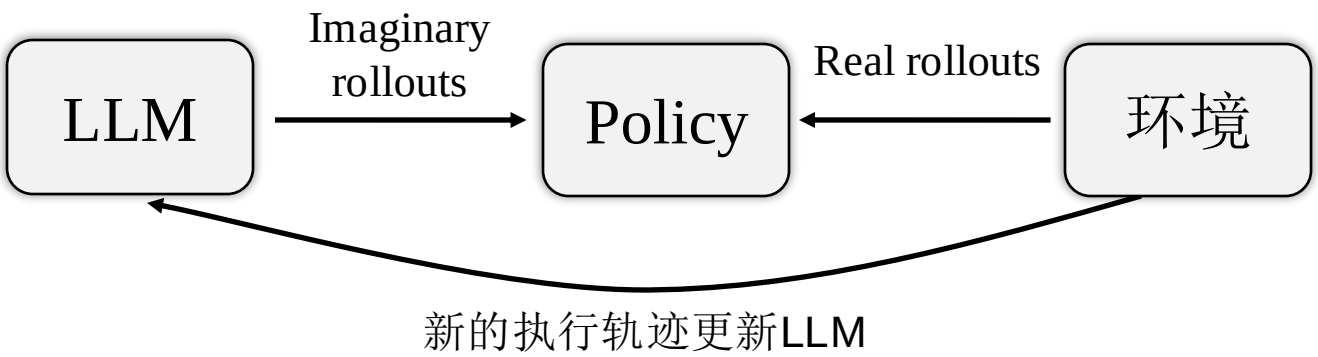
提升决策成功率

极高成功率
泛化性差

提升泛化性

Picture is taken from 'A Survey on Large Language Model based Autonomous Agents. Lei Wang, et al. 2024.'

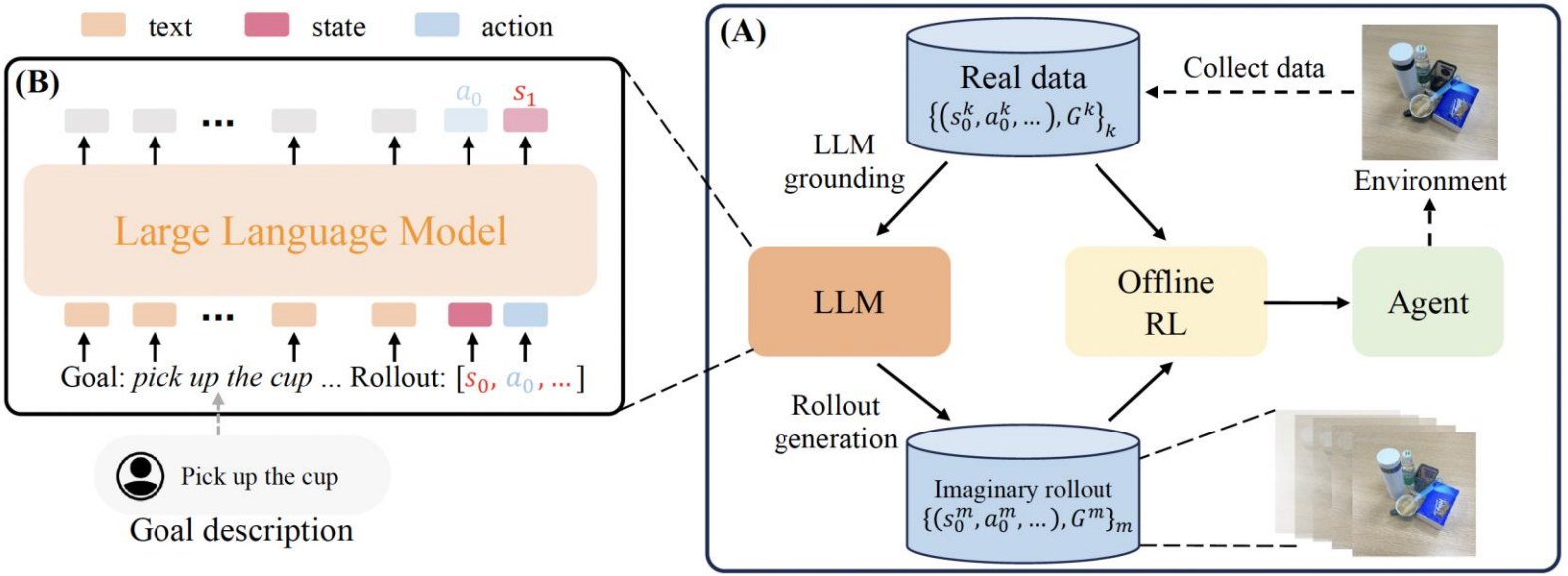
KALM: 释放RL的能量



核心想法:

- LLM不用于规划/拆解，而用于**想象未见任务执行轨迹**
- LLM和执行器都可以从环境反馈中得到优化

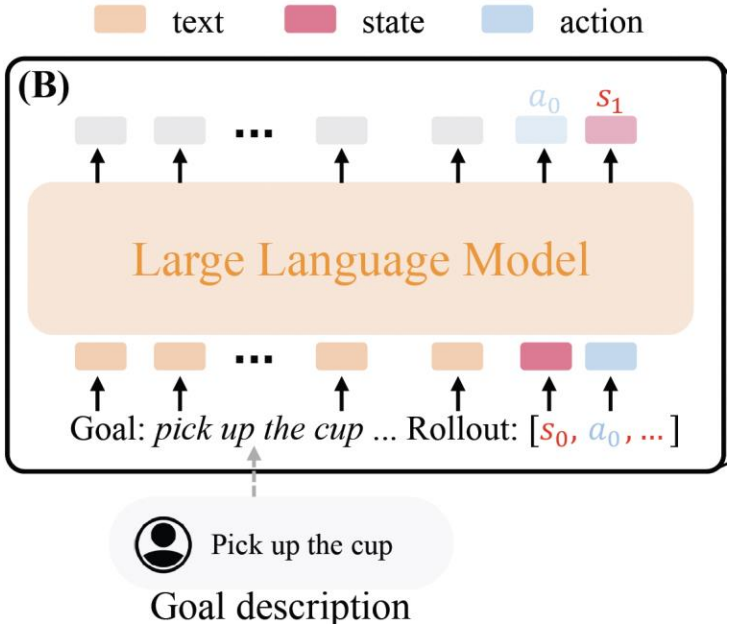
KALM: 释放RL的能量

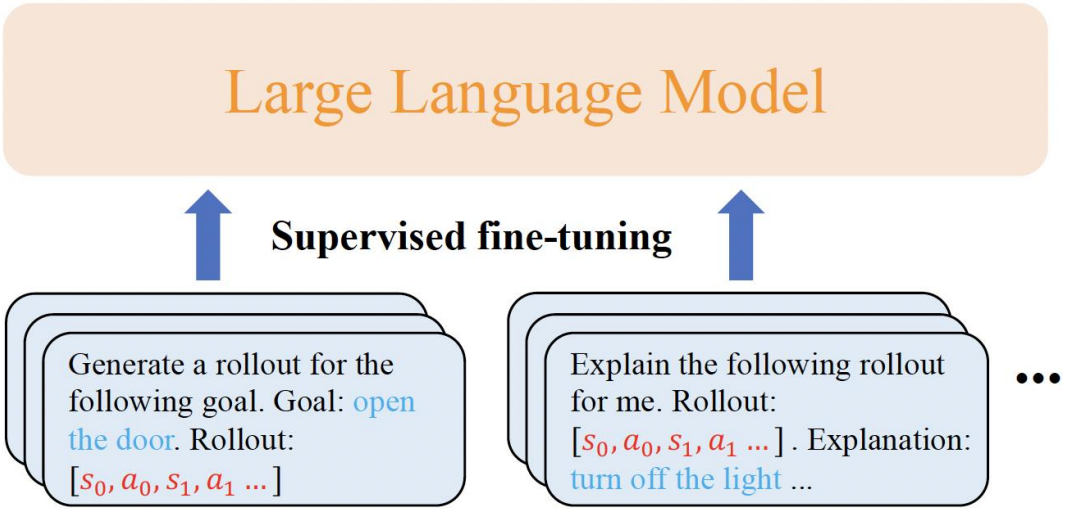


KALM steps:

1. **LLM grounding** that grounds LLM in the control data;
2. **Rollout generation** that generates imaginary rollouts;
3. **Skill acquisition** that trains the policy with RL algorithms.

KALM: 释放RL的能量

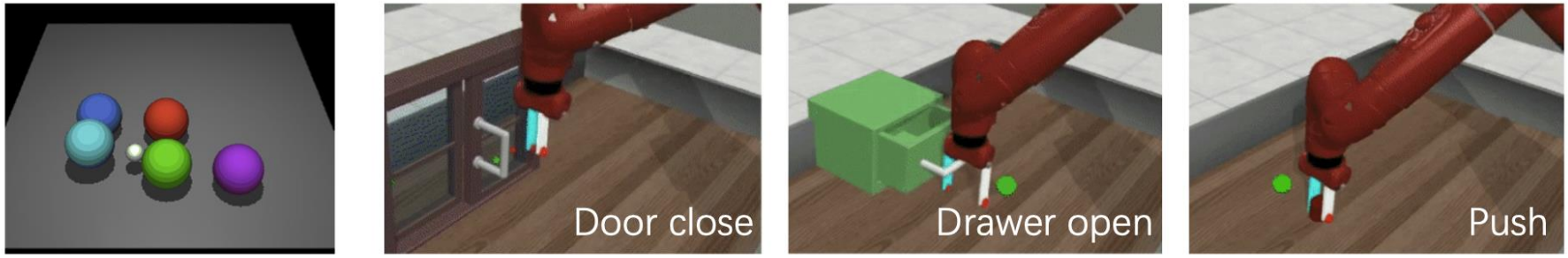




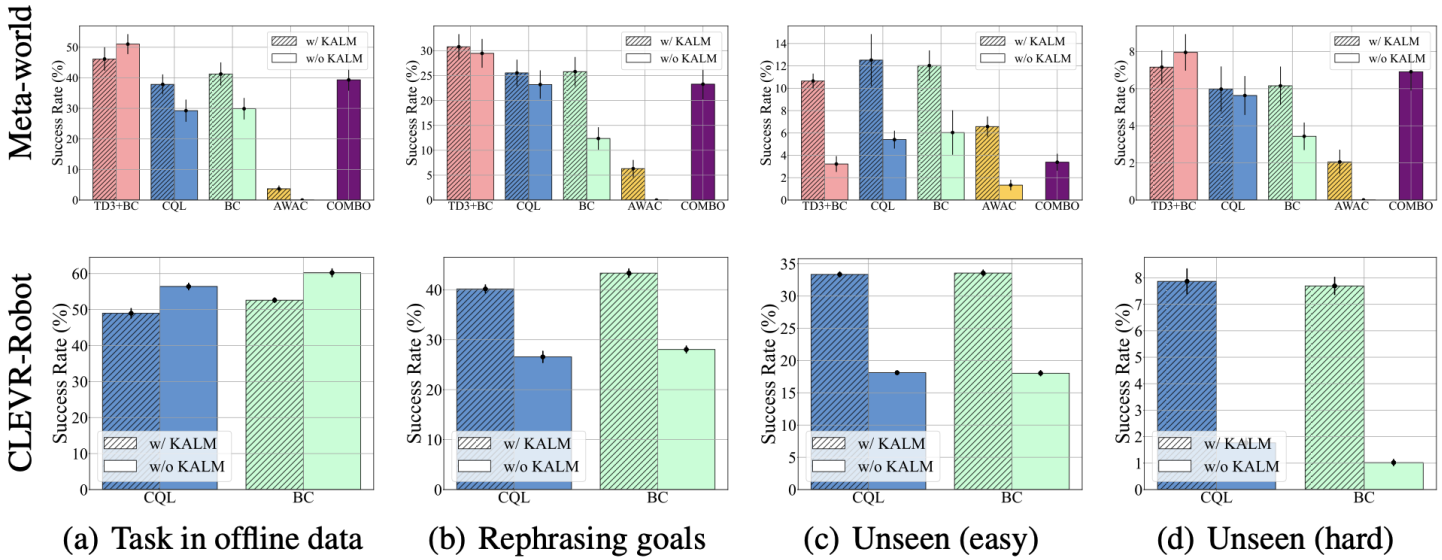
Modify the LLM structure

SFT to understand state, action and dynamics

KALM: 释放RL的能量



(A) CLEVR-Robot (B) Meta-world



(a) Task in offline data (b) Rephrasing goals (c) Unseen (easy) (d) Unseen (hard)

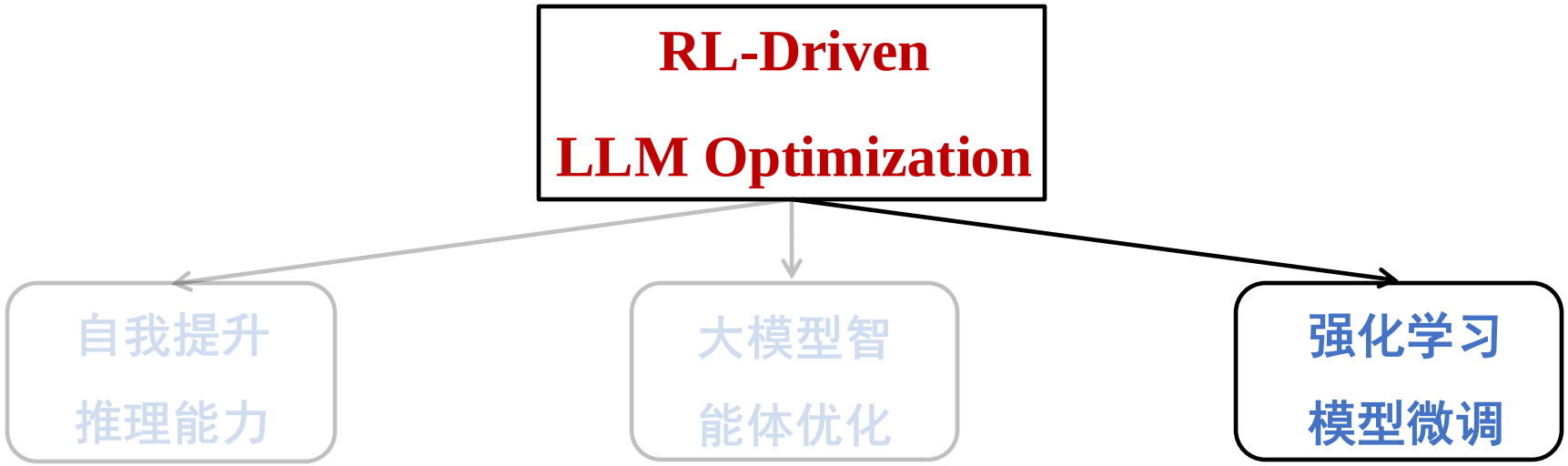
提出方案有效提升智能体在未见任务上执行成功率，超过offline RL领域SOTA方法

下一步

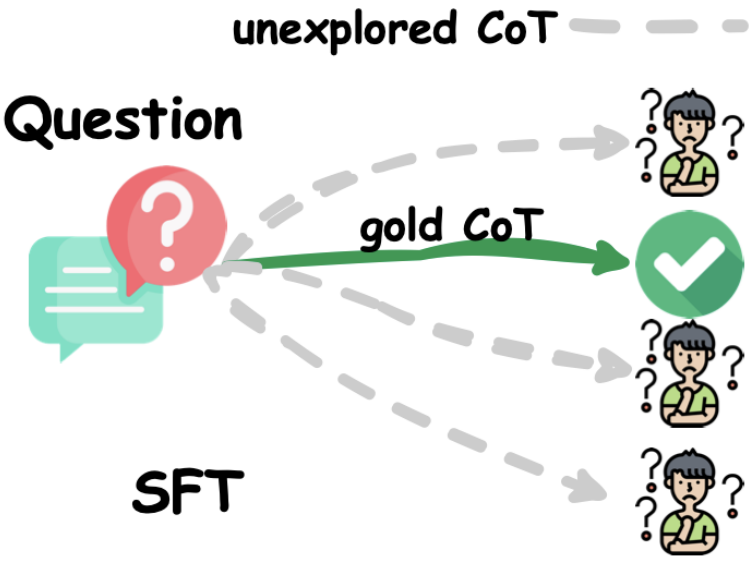
从构建MDP的角度考虑RL优化大模型智能体：

1. Action：如何摆脱直接在海量token空间做决策？
2. Reward：提供有效、稠密的奖赏引导决策；
3. Transition：能否构建世界模型模拟大模型智能体决策路径？
4. State：如何向agent提供有效的任务观测？
5. 。。。

RL驱动的LLM优化前沿问题

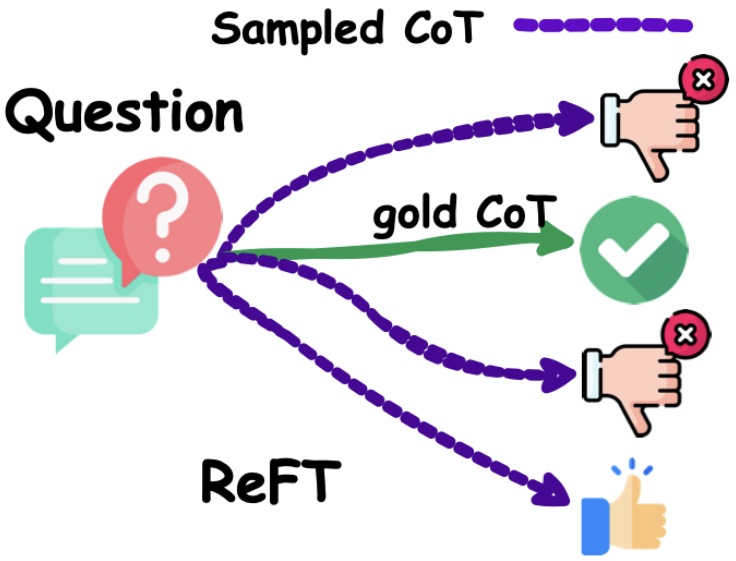


强化学习微调的优势



单一思考路径
泛化性差

监督学习微调



尝试多样思考路径
奖励反馈更精准

强化学习微调

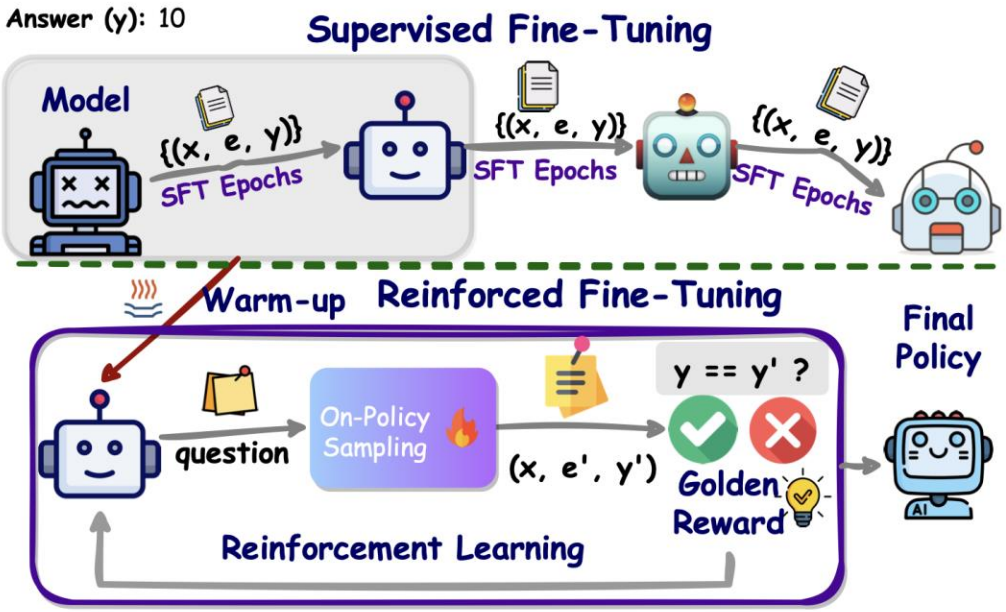
强化学习微调

OpenAI未公开ReFT技术细节，仅字节跳动9月份公开一篇相关文章

Question (x): Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?

Chain-of-Thought (e): We need to calculate her hourly rate and then multiply it by the amount of time she worked. First, we need to convert 50 minutes to hours. There are 60 minutes in an hour, so 50 minutes is equal to 50/60 = 5/6 hours. Next, we can calculate Weng's earnings by multiplying her hourly rate by the amount of time she worked: $12/hour x 5/6 hour = $10. Therefore, Weng earned $10 for 50 minutes of babysitting. The answer is 10.

Answer (y): 10



REFT: Reasoning with REinforced Fine-Tuning. Luong, et al. 2024.

强化学习微调

Algorithm 1: Reinforced Fine-Tuning

Input: $\mathcal{D}_{train} = \{(\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{y})\}$: Tuples of (*question*, *CoT*, *answer*), W: number of warm-up steps, T: number of RL steps, U: number of updates per RL step, $\boldsymbol{\pi}_{\theta}^{(0)}$: Initial policy.

Output: $\boldsymbol{\pi}_{\boldsymbol{\theta}}$: Final policy

$\boldsymbol{\pi}_{\boldsymbol{\theta}} = \boldsymbol{\pi}_{\boldsymbol{\theta}}^{(0)}$

// *Warm-up stage*

for $i \leftarrow 1$ **to** W **do**

  $\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{y} \sim \mathcal{D}_{train}$ // Sample mini-batch from $\mathcal{D}_{train}$

  $\boldsymbol{\theta} = \text{OPTIMIZATION\_STEP}(\mathcal{L}_{SFT}(\boldsymbol{\theta}))$ // Equation 1

// *Reinforcement learning stage*

for $i \leftarrow 1$ **to** T **do**

  $\boldsymbol{x}, \_, \boldsymbol{y} \sim \mathcal{D}_{train}$ // Sample mini-batch without CoT

  $\hat{\boldsymbol{e}} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}(\boldsymbol{x})$ // On-policy CoT sampling

  $\hat{\boldsymbol{y}} \leftarrow \text{EXTRACT}(\hat{\boldsymbol{e}})$ // Extract the answer from CoT

  $\boldsymbol{\pi}_{\boldsymbol{\theta}_{\text{old}}} \leftarrow \boldsymbol{\pi}_{\boldsymbol{\theta}}, V_{\boldsymbol{\phi}_{\text{old}}} \leftarrow V_{\boldsymbol{\phi}}$

  Compute $\delta_t, \hat{A}_t, \hat{R}_t$ using $\pi_{\boldsymbol{\theta}_{\text{old}}}, V_{\boldsymbol{\phi}_{\text{old}}}, \boldsymbol{x}, \hat{\boldsymbol{e}}, \hat{\boldsymbol{y}}$ and $\boldsymbol{y}$

  for $j \leftarrow 1$ **to** U **do**

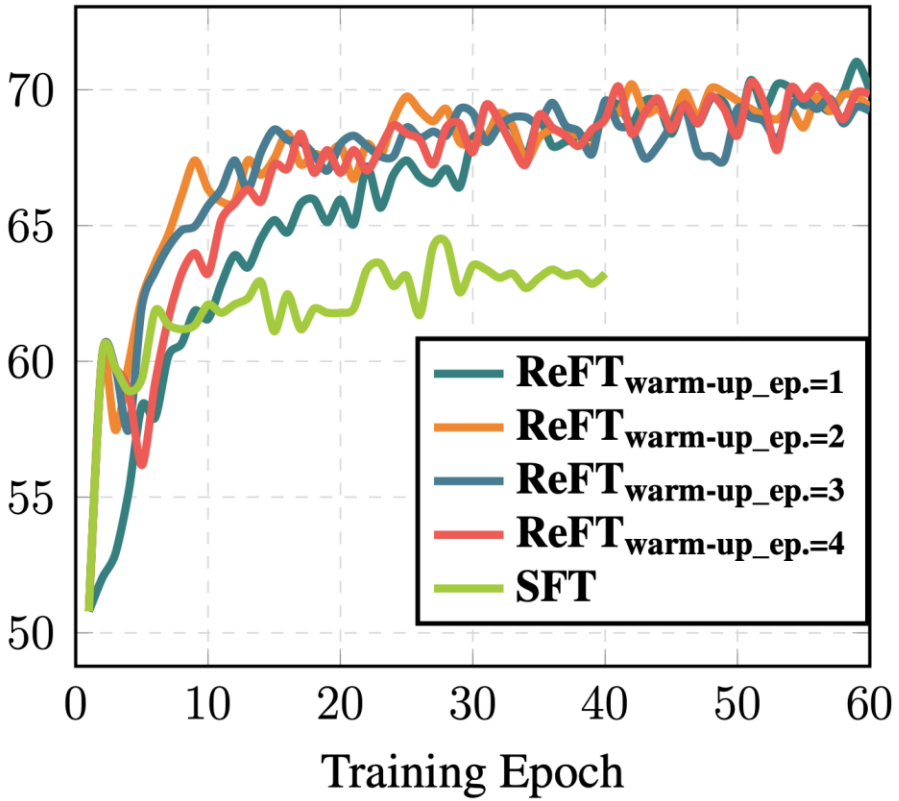
    $\boldsymbol{\theta}, \boldsymbol{\phi} = \text{OPTIMIZATION\_STEP}(\mathcal{L}_{RL}(\boldsymbol{\theta}, \boldsymbol{\phi}))$ // Equation 2

return $\boldsymbol{\pi}_{\boldsymbol{\theta}}$

REFT: Reasoning with REinforced Fine-Tuning. Luong, et al. 2024.

强化学习微调

ReFT训练LLM相比于SFT上限更高



REFT: Reasoning with REinforced Fine-Tuning. Luong, et al. 2024.

参考资料

1. A Simple Overview of the LLM Training Steps. https://weaviate.io/papers/paper20.
2. Understanding the Effects of RLHF on LLM Generalisation and Diversity. Kirk et al. ICLR 2024.
3. https://yaofu.notion.site/GPT-3-5-360081d91ec245f29029d37b54573756. Yao Fu et al.
4. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. Lee, et al. ICML 2024.
5. Language Model Self-improvement by Reinforcement Learning Contemplation. Jing-Cheng Pang, et al. ICLR 2024.
6. Self-Improvement in Language Models: The Sharpening Mechanism. ICLR 2025 submission.
7. Deliberative alignment: reasoning enables safer language models. https://openai.com/index/deliberative-alignment/.
8. LLM Powered Autonomous Agents. Lilian Weng.
9. AutoGPT. https://github.com/Significant-Gravitas/AutoGPT.
10. From Summary to Action: Enhancing Large Language Models for Complex Tasks with Open World APIs. Liu et al.
11. AutoGLM-web. https://ai-bot.cn/autoglm-web/.
12. A Survey on Large Language Model based Autonomous Agents. Lei Wang, et al. 2024.
13. Knowledgeable Agents by Offline Reinforcement Learning from Large Language Model Rollouts. Jing-Cheng Pang, et al. NeurIPS 2024.
14. REFT: Reasoning with REinforced Fine-Tuning. Luong, et al. 2024.

谢谢聆听！

庞竟成
南京大学