



deepseek

DeepSeek-R1

深度解读



ZOMI

视频目录大纲

1. DeepSeek-R1 技术文章解读
2. 从 DeepSeek-R1-Zero 到 DeepSeek-R1
3. Group Relative Policy Optimization(GRPO) 算法原理
4. 对产业的思考与小结



01

DeepSeek-R1 技术文章解读



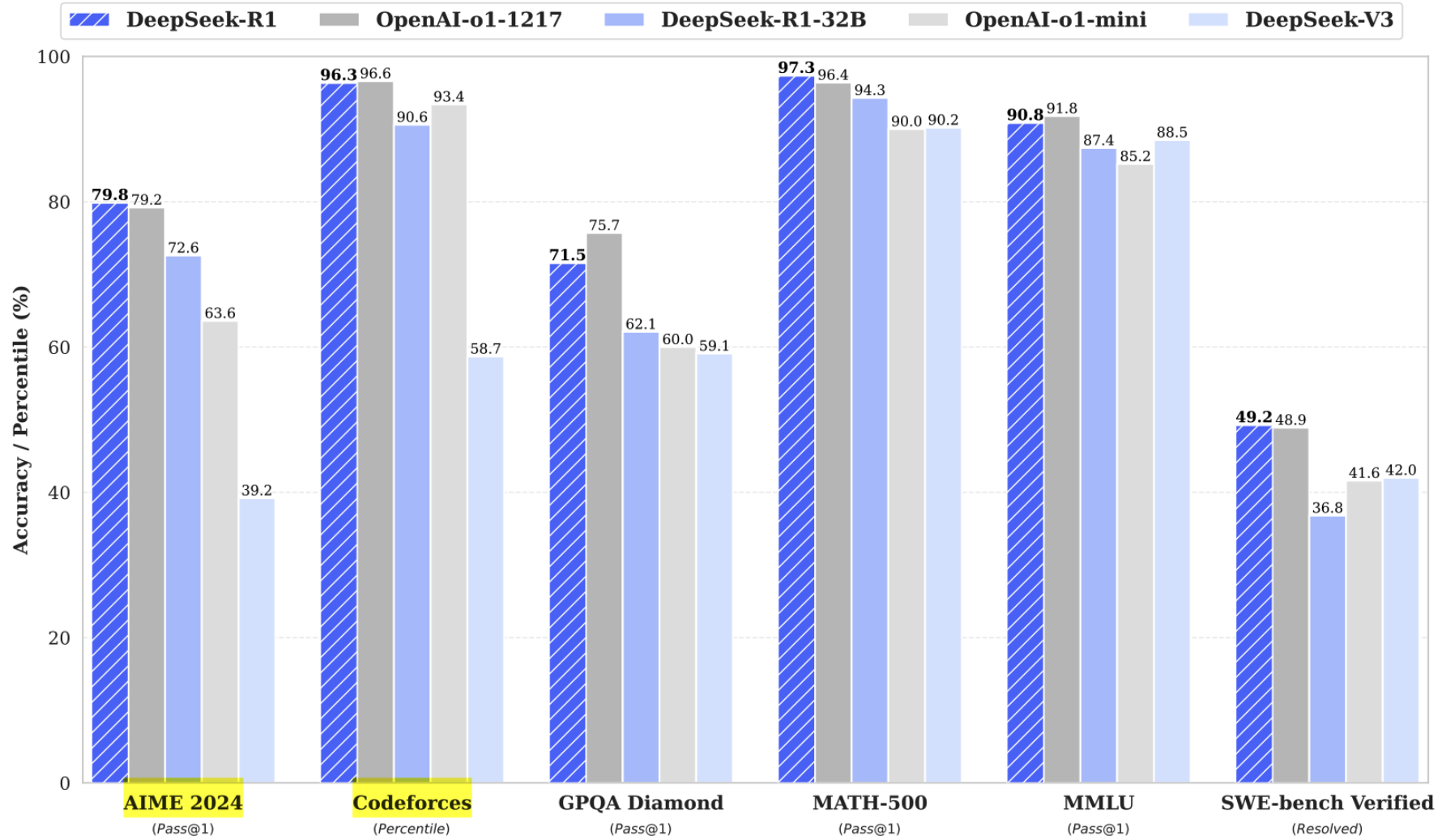


Figure 1 | Benchmark performance of DeepSeek-R1.



摘要

- DeepSeek-R1-Zero是一个通过大规模 强化学习 (RL) 训练出来的模型，没有经过超级微调 (SFT) 这一初步步骤，但却展示了非凡的推理能力。通过强化学习，DeepSeek-R1-Zero 自然而然地出现了许多强大而有趣的推理行为。
- 然而，它也遇到了可读性差和语言混杂等挑战。为了解决这些问题并进一步提高推理性能，我们引入了 DeepSeek-R1，它在RL之前结合了多阶段训练和冷启动数据。
- DeepSeek-R1在推理任务上取得了与OpenAI o1-1217相当的性能。为了支持研究社区，我们开源了DeepSeek-R1-Zero、DeepSeek-R1以及基于Qwen 和Llama从DeepSeek-R1中提炼出的六个密集模型 (1.5B、7B、8B、14B、32B、70B)



Introduciton

- 迈出了利用纯强化学习（RL）提高语言模型推理能力的第一步。我们的目标是探索 LLM 在任何监督数据的情况下开发推理能力的潜力，重点关注它们通过纯强化学习过程进行自我进化的情况。
- 具体来说，我们使用DeepSeek-V3-Base作为基础模型，并采用GRPO（Shao等人，2024年）作为RL框架，以提高模型的推理性能。在训练过程中，DeepSeek-R1-Zero 自然而然地出现了许多强大而有趣的推理行为。
- 经过数千个RL步骤后，DeepSeek-R1-Zero在推理基准测试中表现出了超强的性能。例如，在AIM E 2024上的pass@1得分从15.6%提高到71.0%，在多数投票的情况下，得分进一步提高到 86.7%，与OpenAI-o1-0912的性能不相上下。



DeepSeek-R1-Zero：基础模型的强化学习

- 强化学习在推理任务中表现出了显著的有效性，我们之前的工作（Shao 等人，2024 年；Wang 等人，2023 年）也这一点。证明了然而，这些工作在很大程度上依赖于监督数据，而监督数据的收集需要大量时间。
- 探索 LLMs 在没有任何监督数据的情况下开发推理能力的潜力，重点关注它们通过纯强化学习过程进行自我进化的情况。我们首先简要介绍了我们的强化学习算法，然后介绍了一些令人兴奋的结果，希望能为社区提供有价值的见解。



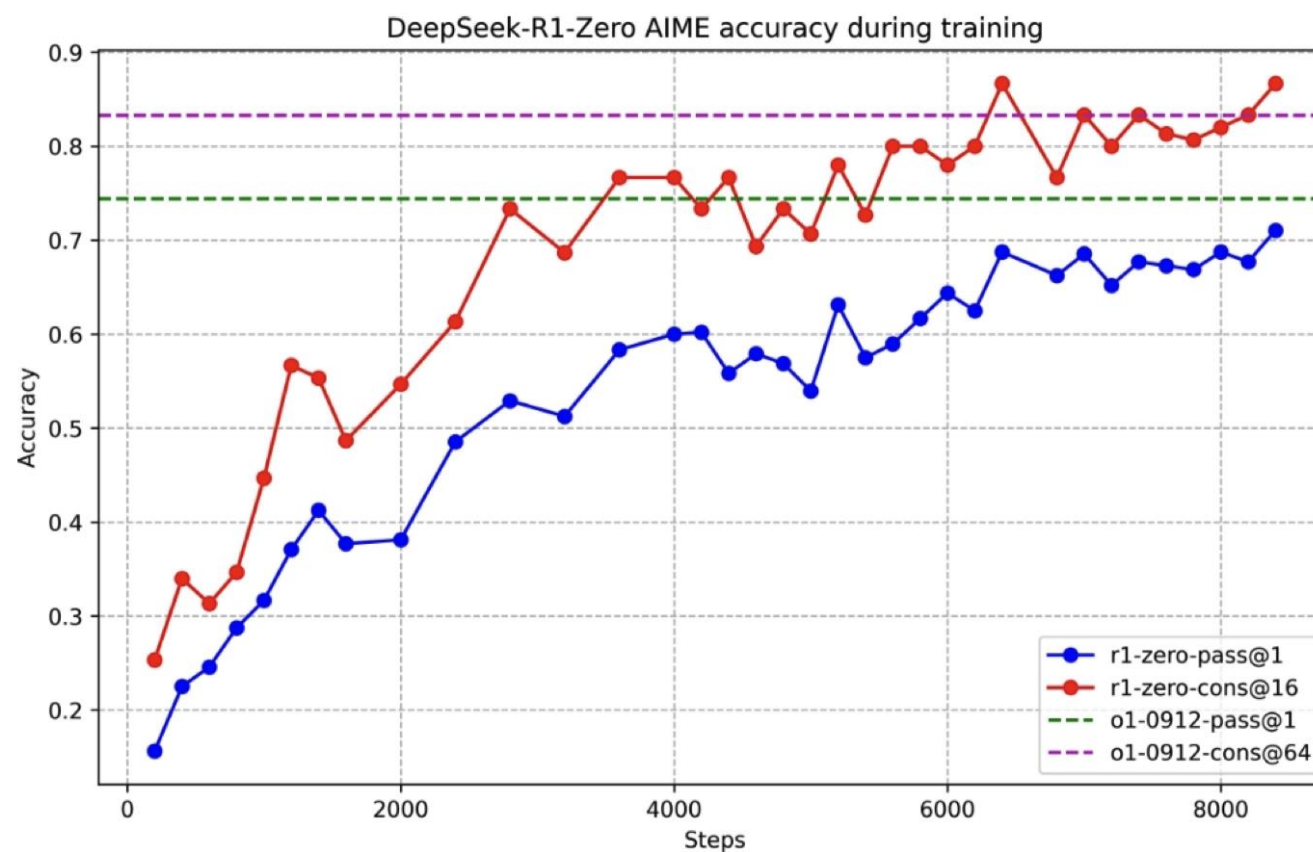
DeepSeek-R1-Zero 的奖励模型

- 奖励是训练信号的来源，它决定了 RL 的优化方向。为了训练 DeepSeek-R1-Zero，我们采用了基于规则的奖励系统，主要包括两类奖励：
 - 准确性奖励：准确性奖励模型可评估答案是否正确。对于具有确定结果的数学问题，模型需要以指定格式（如在方框内）提供最终答案，从而对正确性进行可靠的基于规则的验证。同样，对于 LeetCode 问题，可以使用编译器根据预定义的测试用例生成反馈。
 - 格式奖励：除了准确性奖励模型外，我们还采用了格式奖励模型，强制模型将其思考过程置于"< 思考>"和"</思考>"标记之间。



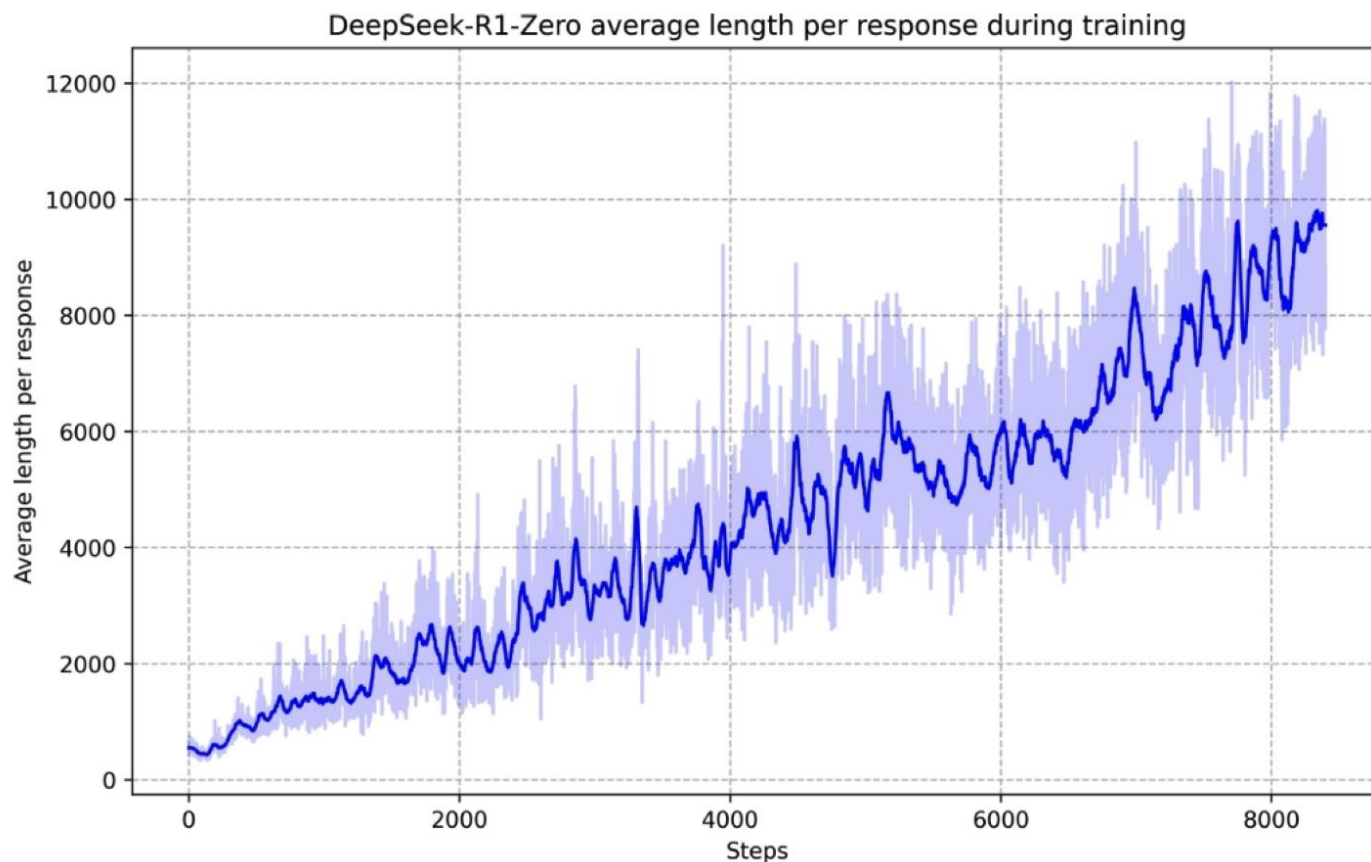
DeepSeek-R1-Zero 的性能

- DeepSeek-R1-Zero 在训练过程中的 AIME 准确率。我们对每个问题抽取 16 个回答，计算总体平均准确率，以确保评估的稳定性。



DeepSeek-R1-Zero的自我进化过程

- 在推理过程中，DeepSeek-R1-Zero 在训练集上的平均响应长度。DeepSeek-R1-Zero 自然而然地学会了用更多的思考时间来解决推理任务。



DeepSeek-R1-Zero 的 "啊哈时刻"

- 无需明确教导模型如何解决问题，只需向它提供正确的激励，它就能自主开发出先进的问题解决策略。

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ...

$$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...



DeepSeek-R1：冷启动强化学习

- 1) 作为冷启动，纳入少量高质量数据能否进一步提高推理性能或加速收敛？
- 2) 如何才能训练出一个用户友好型模型，不仅能生成清晰连贯的思维链（CoT），还能展示强大的通用能力？



DeepSeek-R1：冷启动

- 构建并收集了少量长CoT数据，以微调模型作为初始RL行为体。为了收集这些数据，我们探索了几种方法：
 - 以长 CoT 为例，使用少量提示；
 - 直接提示模型生成详细答案，并进行反思和验证；
 - 以可读格式收集 DeepSeek-R1- Zero 的输出结果；通过人工注释者的后期处理完善 结果。



DeepSeek-R1：以推理为导向的强化学习

- 在 RL 训练过程中引入了语言一致性奖励，其计算方法是 CoT 中目标语言单词所占的比例。
- 虽然消融实验表明这种一致性会导致模型性能略有下降，但这种奖励符合人类的偏好，使其更具可读性。
- 最后，我们将推理任务的准确性和语言一致性奖励直接相加形成最终奖励。然后，我们对微调后的模型进行强化学习（RL）训练，直到它在推理任务上达到收敛为止。



DeepSeek-R1：拒绝采样和监督微调

- 通过纳入更多数据来扩展数据集，其中一些数据使用了生成式奖励模型，将地面实况和模型预测输入 DeepSeekV3 进行判断。
- 此外，由于模型输出有时混乱难读，我们过滤掉了混合语言的思维链、长段落和代码块。对于每个提示，我们都会对多个回答进行抽样，只保留正确回答。



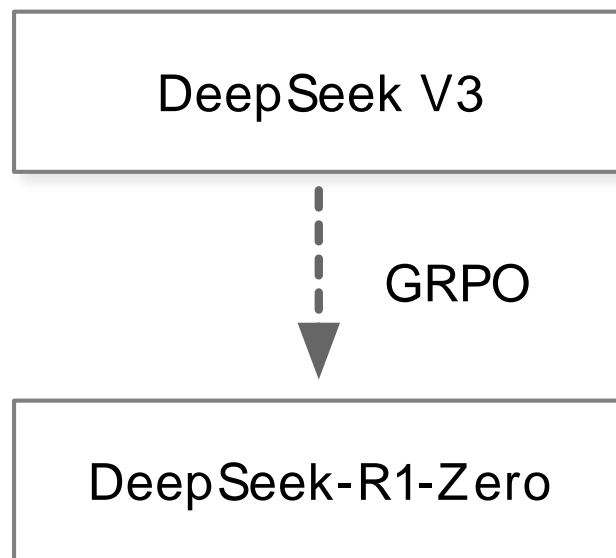
02

DeepSeek-R1-Zero

DeepSeek-R1



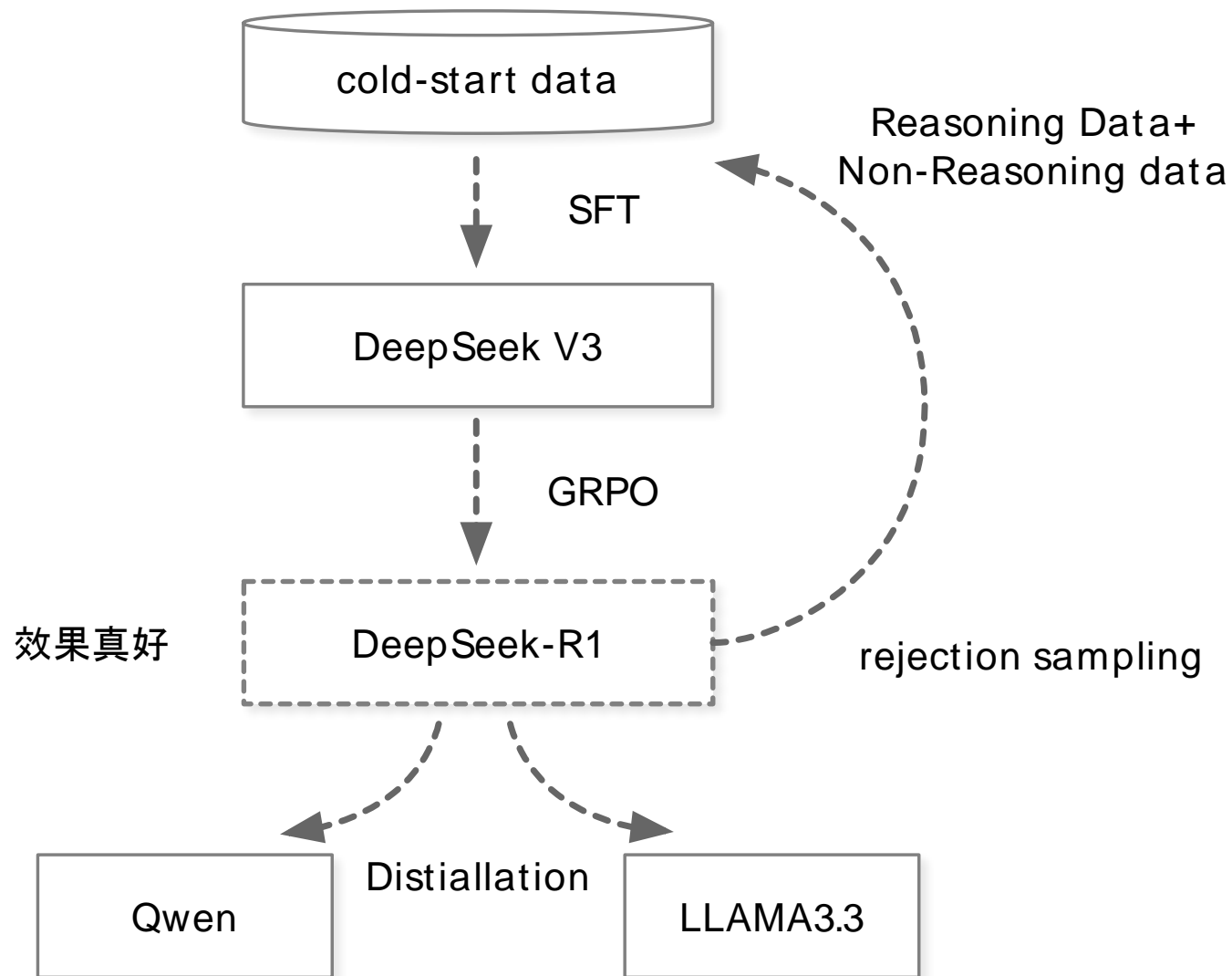
DeepSeek-R1-Zero 整体训练流程



Good: 自我进化能力, 初步具备 Test-Time Reasoning

Bad: Reasoning 过程可读性差, 中英文混淆

DeepSeek-R1 整体训练流程



03

GRPO 算法原理



- 最早 GRPO 工作发表于 DeepseekMath , 是数学逻辑 reasoning 上使用的, 为了了解细节, 建议还是先阅读 DeepseekMath 论文。

DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

Zhihong Shao^{1,2*†}, Peiyi Wang^{1,3*†}, Qihao Zhu^{1,3*†}, Runxin Xu¹, Junxiao Song¹
Xiao Bi¹, Haowei Zhang¹, Mingchuan Zhang¹, Y.K. Li¹, Y. Wu¹, Daya Guo^{1*}

¹DeepSeek-AI, ²Tsinghua University, ³Peking University

{zhihongshao, wangpeiyi, zhuqh, guoday}@deepseek.com

<https://github.com/deepseek-ai/DeepSeek-Math>

300 年前 鹅也是个研究算法的人



深度强化学习原理与实践(异步图书出品)

陈仲铭 何明

本书构建了一个完整的深度强化学习理论和实践体系：从马尔可夫决策过程开始，根据价值函数、策略函数求解贝尔曼方程，到利用深度学习模拟价值网络和策略网络。书中详细介绍了深度强化学习相关算法，如Rainbow、Ape-X算法等，并阐述了相关算法的具体实现方式和代表性应用（如AlphaGo）。此外，本书还深度剖析了强化学习各算法之间的联系，有助于读者举一反三 更多

RL + LLM 知识

- LLM 中主流 RLHF 方向分为两大路线:
 1. 以 [PPO] 为代表的 On Policy 路线
 2. 以 [DPO] 为代表的 Off Policy 路线

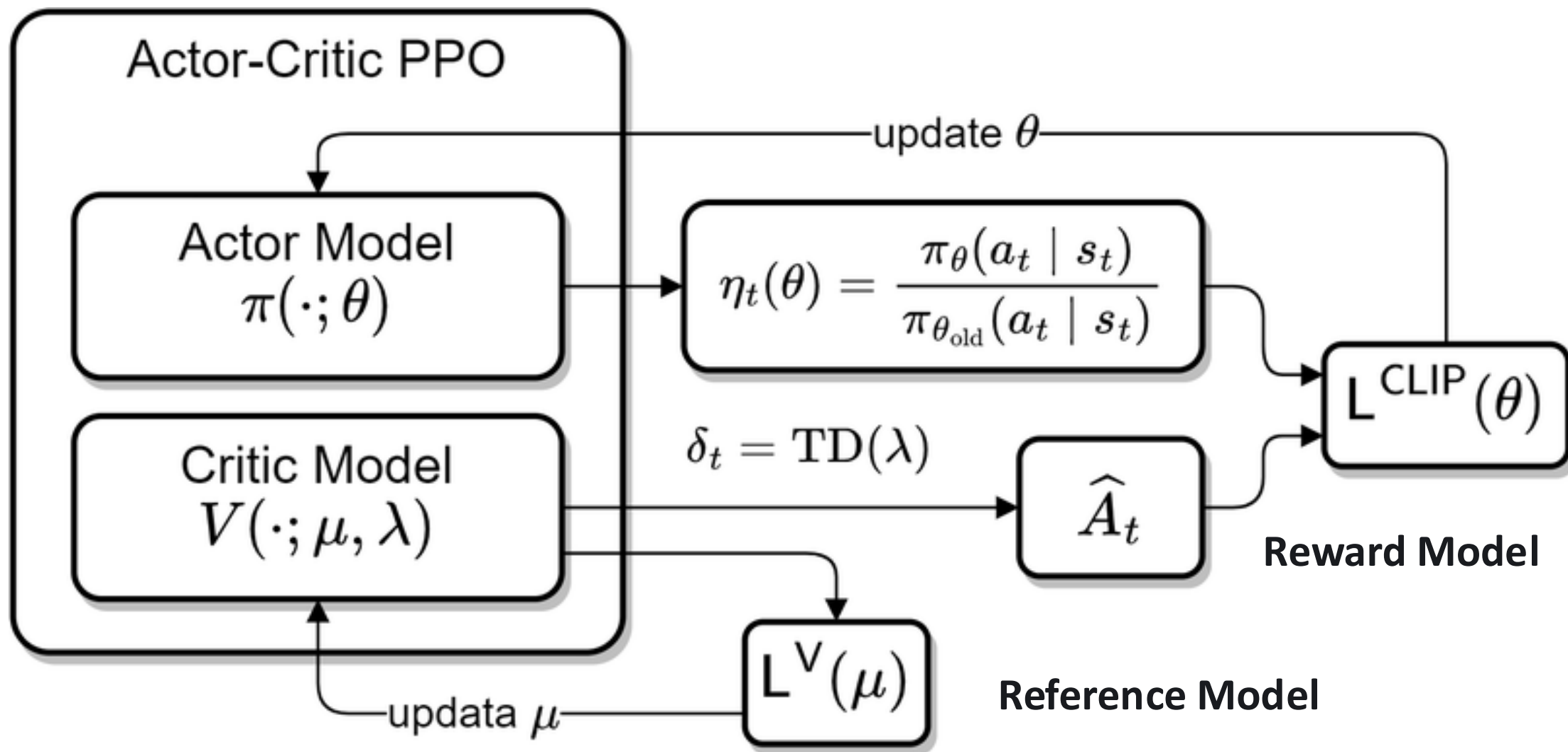


RL + LLM 知识：On Policy 与 Off Policy 差异在哪？

- On policy:
 - 每次训练都基于自己的生成模型（Actor），通过教练（Critic）反馈奖励；
 - 优势是效率高，没有模型自生成自然效率高，问题是训练后模型能力可能不够；
- off policy:
 - 基于现有标注的情况进行分析，存在训练样本可能与模型不匹配的问题；
 - 优势是更有可能达到模型能力的上限，问题是效率较低。



PPO 算法

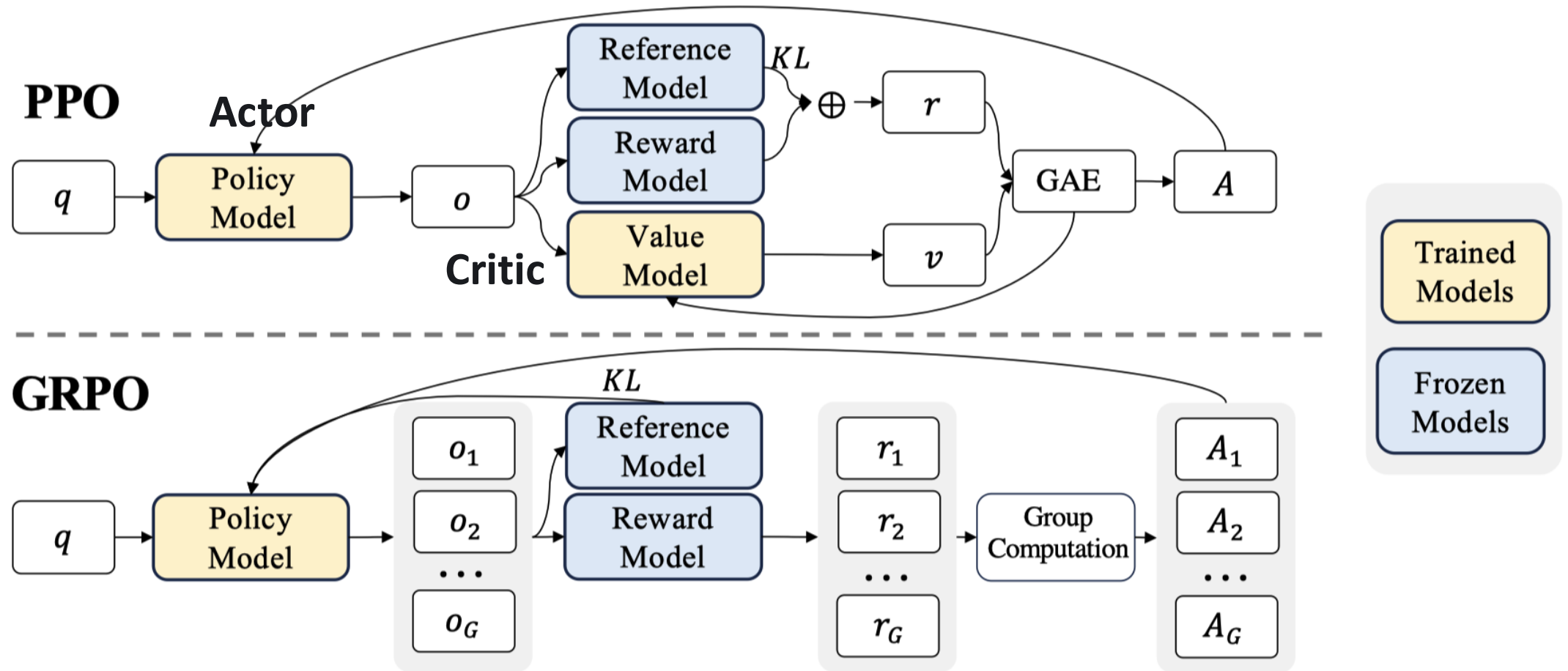


PPO 问题

- 从PPO的优化过程分析，其存在如下缺点：
 1. 需要训练一个与策略模型大小相当的价值模型，这带来了巨大的内存和计算负担；
 2. LLM 通常只有最后一个 token 会被奖励模型打分，训练在每个 token 上都准确价值函数难；
- 1. GRPO 避免了像 PPO 那样使用额外的 Value Model 近似，而是使用同一问题下多个采样输出的平均奖励作为基线



Group Relative Policy Optimization (GRPO)



Group Relative Policy Optimization (GRPO)

- **无需额外的价值函数：**
 - GRPO 使用组内平均奖励作为基线，避免了训练额外的价值函数，从而减少了内存和计算负担。
- **与奖励模型的比较性质对齐：**
 - GRPO 使用组内相对奖励计算优势函数，这与奖励模型通常在同一问题的不同输出之间进行比较的性质相符。
- **KL惩罚在损失函数中：**
 - GRPO 直接将训练策略 π_θ 和参考策略 π_{ref} 之间的 KL 散度添加到损失中，而不是像 PPO 那样在奖励中添加 KL 惩罚项，从而避免了复杂化 $A^{i,t}$ 的计算。

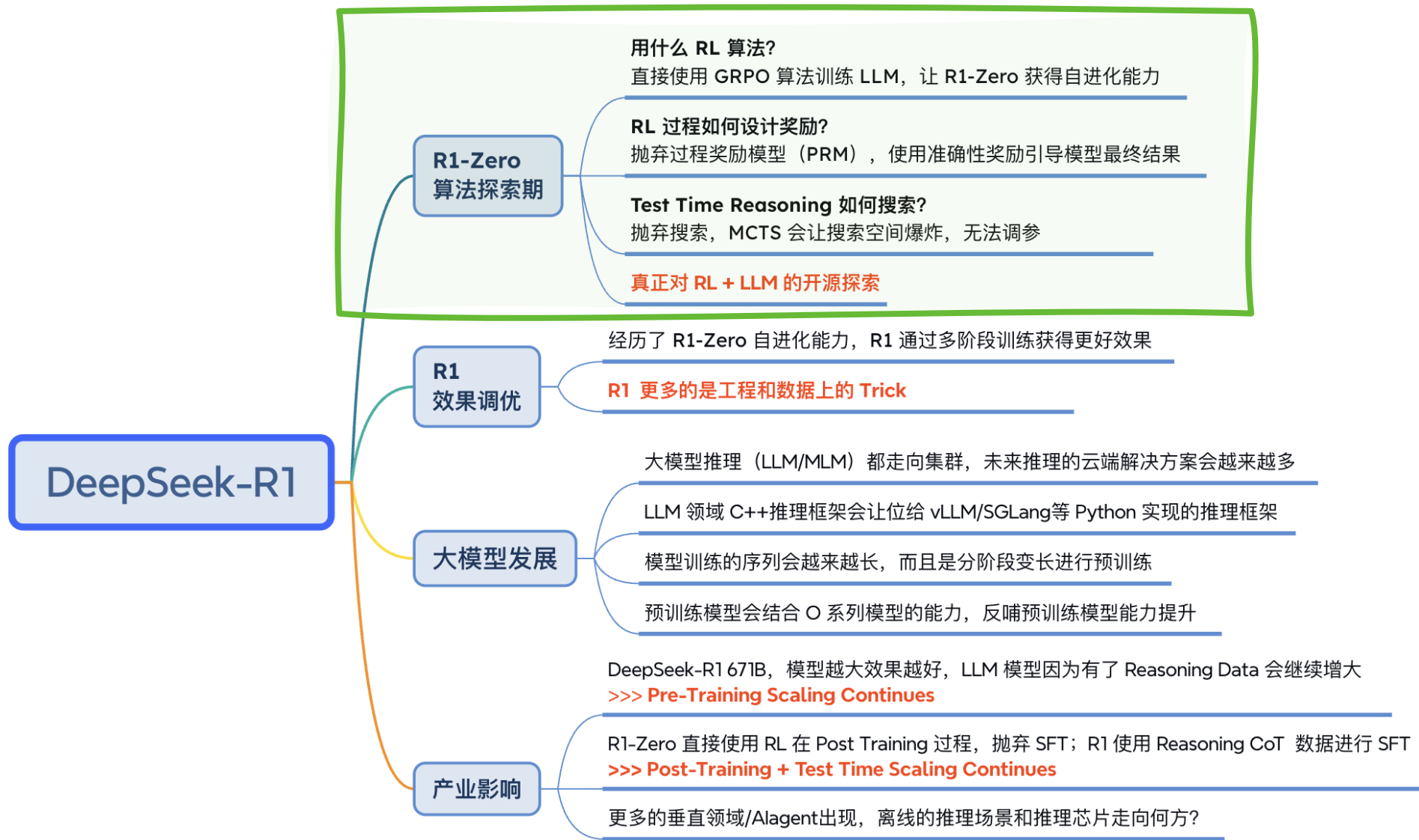


04

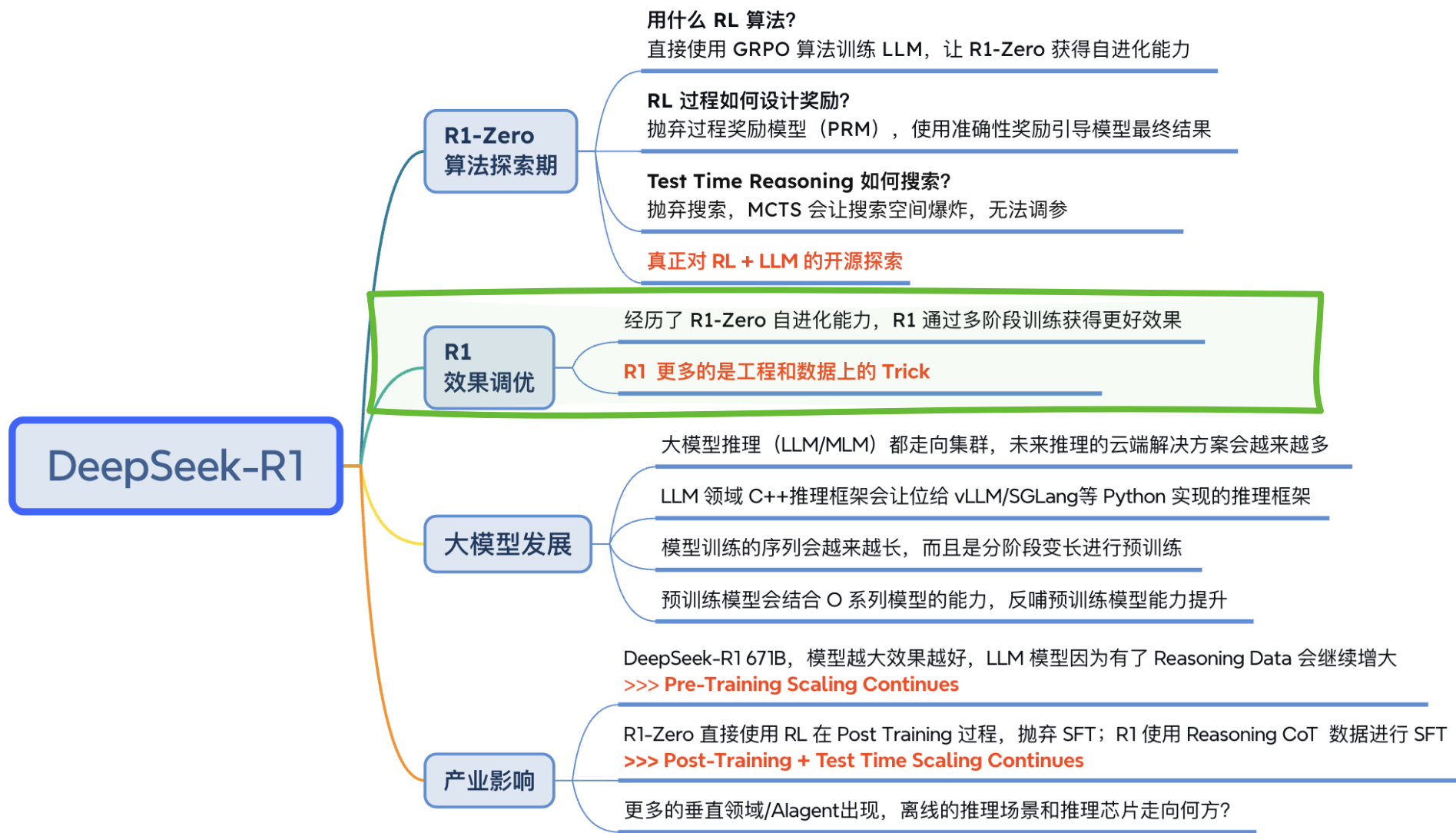
对产业的思考与小结



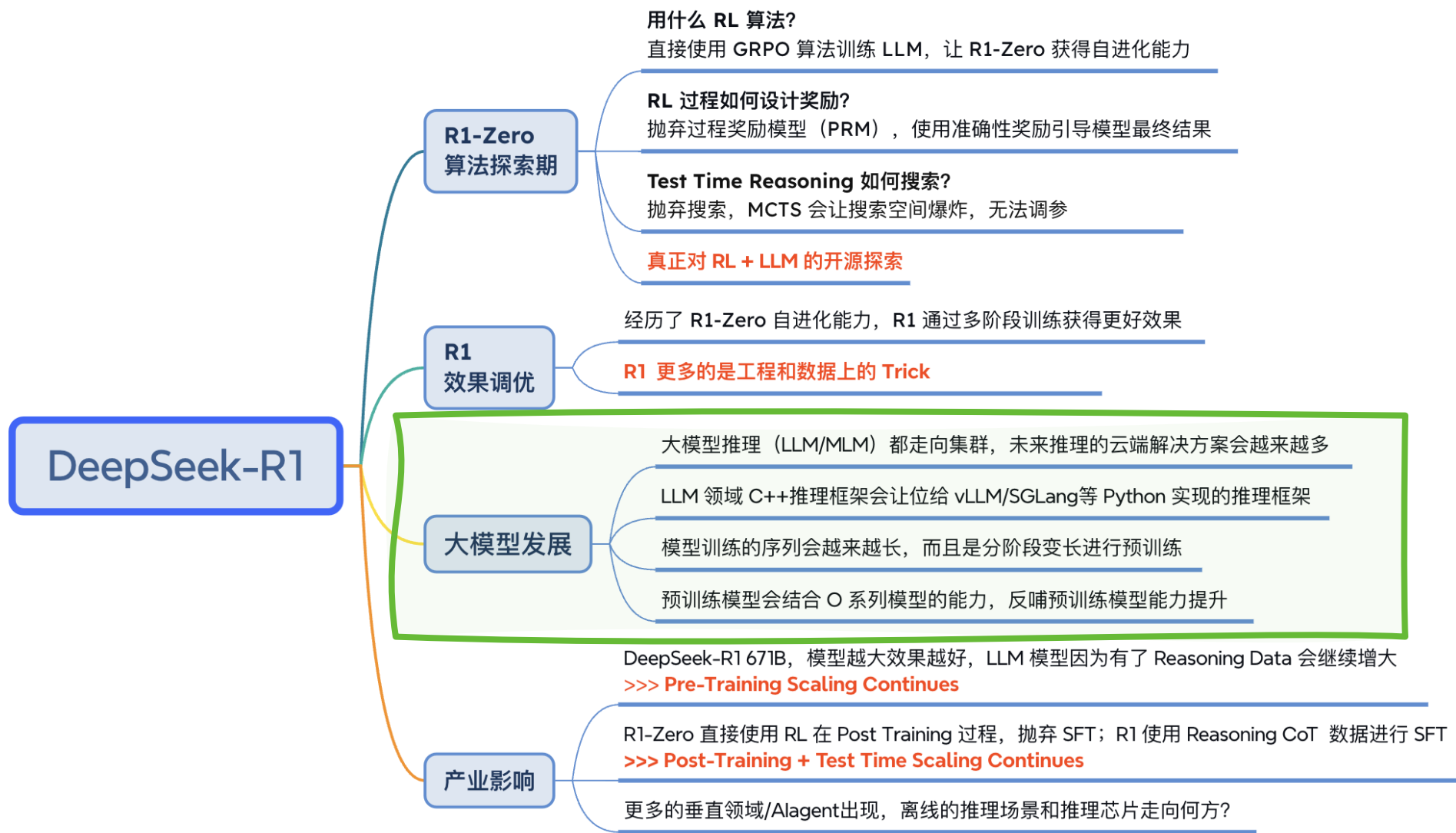
总结与思考



总结与思考



总结与思考



总结与思考





Thank you

把AI系统带入每个开发者、每个家庭、
每个组织，构建万物互联的智能世界

Bring AI System to every person, home and
organization for a fully connected,
intelligent world.

Copyright © 2024 XXX Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.



GitHub <https://github.com/chenzomi12/AIFoundation>

引用与参考

- <https://huggingface.co/deepseek-ai>
- <https://mp.weixin.qq.com/s/atKyfC5l-BaStje8-F3FGQ>
- https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf
- PPT 开源在:
- <https://github.com/chenzomi12/AIFoundation/tree/main/09News/00Others>



留给读者思考

1. DeepSeek-R1 如何通过引入冷启动数据来提高推理性能？这个方法是否适用于多模态大模型？
2. DeepSeek-R1 相比之前的版本在哪些方面有所改进？它在STEM相关问题上的表现为什么更好？
3. 文章提到 DeepSeek-R1 在某些任务上表现不如 DeepSeek-V3，这是为什么？有没有改进的方案？

