

Deepseek 分析

理解 Deepseek

Deepseek R1 底层技术带来的两点变化

后训练+推理算法层面的变化，R1-zero 替代传统 LLM 推理能力，不再用 SFT、MCTS 等复杂例证+奖励方式规范模型思考路径，直接选择强化学习，给予 LLM 准确性与格式奖励，在不断尝试获得反馈的过程中，R1-zero 顿悟。过程中强化学习的方式是 GRPO，群体相对政策优化，同个问题多种回答进行自我采样+比较，就像老师出了一道题，多个同学同时回答，只用高分逻辑找到最佳答案路径更新模型，最后顿悟

论文中提到，Deepseek 观察到的 R1-zero 顿悟 moment，看到某数学公式后提到 “Wait, wait. Wait. That's an aha moment I can flag here” （等等、等等、这是个值得标记的啊哈时刻），这在传统 SFT 模型中不存在

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

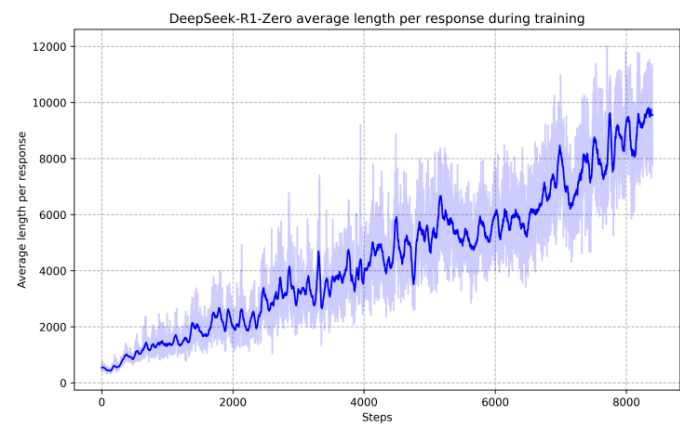
Response: <think>
To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ...
 $(\sqrt{a - \sqrt{a + x}})^2 = x^2 \implies a - \sqrt{a + x} = x^2$.
Rearrange to isolate the inner square root term:
 $(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
...
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: ...
...

(R1-zero 自我发现 aha moment)

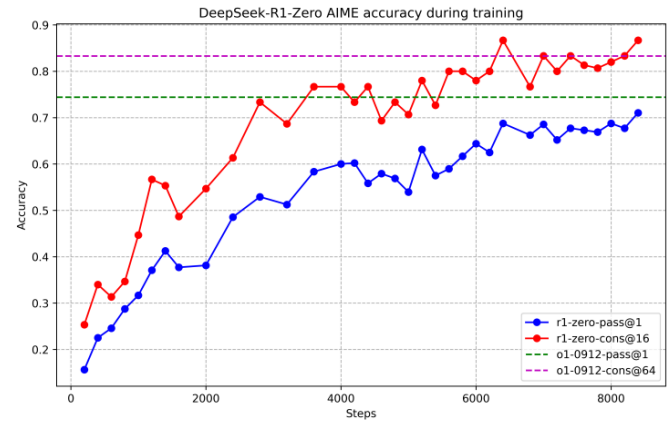
除此以外，R1-zero 在长期自我采样+比较的推理过程中，学会了控制推理时间，Deepseek 发现 R1-zero 的响应时间越久，平均步骤越多（跳跃点），进步越快（数据集准确率），长期而言 R1-zero 具

有了对待不同深度问题，分配算力与步骤的能力，找到最“投机取巧”解决方式，摆脱人的思维模式，形成最自然的推理

在全球最具深度创造性解题测试集，AIME 竞赛中，Deepseek 团队观测到 R1-zero 能够根据不同问题复杂度自动调节响应长度，其正确率从 15.6% 逐渐攀升至 71.0%，证明其不是在简单套用模版，而在理解问题



（推理时间 vs “跳跃点” 步骤）



（“跳跃点” 步骤 vs AIME准确率）

R1-zero 像天才儿童自创的解题方法不可读，因此在 R1-zero 基础上引入冷启动数据和多阶段训练，将人类的思维表达方式微调到 R1-zero 模型中，让其思维过程具有可读性，这就是 R1

请告诉我 deepseek-R1 的技术细节与原理

🔍 已搜索到 39 个网页 >
🧠 已深度思考 (用时 12 秒) ^

嗯，用户问的是关于DeepSeek-R1的技术细节和原理。我需要仔细分析提供的搜索结果，提取关键信息。首先，DeepSeek-R1是一个专注于推理的AI模型，对吧？根据网页1和网页3的内容，它主要采用了强化学习（RL）技术，没有依赖传统的监督微调（SFT）。网页3提到他们用了Group Relative Policy Optimization（GRPO）算法，这可能是一个关键点。

然后，训练流程方面，网页1和网页3都提到了分阶段的训练方法。比如，DeepSeek-R1-Zero是完全通过RL训练的，而R1则结合了冷启动数据和多阶段RL。冷启动阶段可能涉及引入一些高质量的推理链数据来规范输出格式，这对解决可读性问题有帮助。另外，网页7提到他们使用了两阶段的强化学习，一个是推理导向的，另一个是对齐人类偏好。

（R1 的思考过程可读性）

训练层面的算法+工程创新，R1 的成功离不开底层基模 V3 的低成本训练，V3 之所以成本相较于 GPT、Llama 系列低这么多，是因为 MLA 机制、FP8 乘法、DualPipe 并行训练算法等

1. Deepseek-R1 以三十分之一的 API 成本达到 GPT-o1 级别的表现

- GPT-o1 output=\$60.00 / 1M tokens, input=\$15.00 / 1M tokens
- Deepseek-R1 output=\$2.19 / 1M tokens, input=\$0.55 / 1M tokens (cache miss)

2. Deepseek-R1 以十分之一的训练成本达到 Llama 3.1 级别的表现

- V3 利用2048个H800训练成功，每万亿 token 仅需180K个H800 GPU小时，总计278万GPU小时的训练成本，Llama 3.1训练使用16,384块H100，总计2100多万GPU小时，翻了十倍

MLA，过去 Transformer 每一层都有计算+存储，Key+Value矩阵占用大量内存，MLA 在Transformer 架构中引入动态层聚合机制，通过动态合并相邻层特征减少计算量

FP8，训练矩阵乘法以FP8，8位浮点数进行，相比FP32和FP16，满足最低精度要求，占用空间小，计算快

并行，利用DualPipe流水线并行训练，专家模块分配到不同服务器训练，同时让通信与计算重叠，提升计算效率



Figure 5 | Example DualPipe scheduling for 8 PP ranks and 20 micro-batches in two directions. The micro-batches in the reverse direction are symmetric to those in the forward direction, so we omit their batch ID for illustration simplicity. Two cells enclosed by a shared black border have mutually overlapped computation and communication.

(8个流水线并行的 DualPipe 调度示例)

除此以外，Deepseek 还结合了多个工程技术，包括无辅助损失的负载均衡策略，根据专家历史利用率动态调整“接收容量”，底层通信优化，跨节点全对全通信内核，MTP技术，多token预测，蒸馏，沉淀复杂模型数据..

这些技术来自于 Deepseek 过去在量化领域积累的大量庞大集群服务器调教经验（15年百卡，19年千卡，21年万卡），例如 Deepseek 在训练架构中手撸了特殊算子，除英伟达之外还能支持 AMD，此类技术数不胜数

Deepseek 带来了什么样的创新？

工程组织能力的向下 Scaling Law，当前 LLM 的双性能曲线，向上是更深层次预训练、推理，向下是效率提升与落地能力，R1 在向下演化的突破让行业更惊艳，当前模型难以落地很大程度因为成本还不够低（强化学习时代同样要求更高推理成本），Deepseek 组合出了一整套已知原理的最低成本实践，MLA 多层注意力存在已久，MTP 来自24年4月 Meta 的论文（应用的比 Meta 更快），FP8 技术来自英伟达，配合 Deepseek 底层的负载均衡、DualPipe 工程能力，清晰的把握每个技术最大边界，组合降本

成本是加快技术落地不可或缺的技术创新，第一次工业革命有一个非常著名的经济悖论，Jevons 悖论，当时人们普遍认为蒸汽机越来越节能，越来越会减少煤炭消耗，但实际情况却打了所有人的脸，更高效的蒸汽机降低了使用煤炭的成本，导致蒸汽机被更广泛应用，煤炭消耗总量反而大幅增加，Deepseek 提供了更高效的蒸汽机

虽不是完全开源，但比闭源模型提供了太多可参考细节，开源在23年 Llama 2 被 Meta 初次定义，仅开放模型权重+推理代码，不开源训练代码，从此之后 Qwen、Gemma、百川都遵循了这样的开源格式。很多人说 Deepseek 没有真正开源，但它其实没有特例独行，遵循了 LLM 的开源惯例，且用了比 Apache 更开放的 MIT license

训练代码在于用户知道如何将模型“炼出来”，中间有很多工程上的小技巧并未开放，推理代码能够让用户知道怎么用模型，能够为更多本地部署、微调模型的专业用户，以及社区贡献者带来更好的模型提供思路

举个例子，R1 开源对高校团队的帮助，让团队更了解哪些技术 work 哪些不 work，例如混合专家、大模型小型化蒸馏、量化压缩、MLA 等，了解部署细节，快速试错，在有限 GPU 里尝试更落地的方案

Deepseek 带来了什么样的投资机会？

安全，R1 发布第一周，美国除了呼吁禁卡，声称抄袭，第三个重要行为是攻击，大量攻击IP来源于美国，Deepseek 1月30那天较2天前攻击指令暴增百倍，攻击范围从 SSDP、NTP 反射等初始攻击，扩展到大量 HTTP 代理攻击，再到僵尸网络（botnet），面临的攻击手段越多，防范难度越大，后 Deepseek 在一周内两次宕机

研究类问答类 Agent，在 R1 发布的一周之内，OpenAI 紧急发布了 o3-mini 和 Deep Research 两个版本模型，o3-mini 作为高级推理模型，已公开思维链，正在考虑开源。Deep Research 在o3 的推理能力之上加上联网搜索，同时能够自动执行 Python 任务，事先会与用户进行细节确认，然后进行全流程的发现、推理（浏览图片、表格、PDF后，确定每一步做什么）、综合，研报生成体验是信息全面、绩

密，但稳定性较差，指令要求高，仍预示着金融、科技、政策、工程领域的决策辅助类 or 研究型 Agent 兴起

中文/各国自主语言 LLM，Deepseek 展现出强大的网文感，中文优美感，春节期间小红书上大量网友用 R1 写诗歌、网文、小说，跳脱性、逻辑性、文字底蕴型超出 o1 等推理模型，甚至包括很多中文梗，这部分能力来自 R1 在其他领域形成的强逻辑能力的二次泛化，同时也受益于团队亲自下场寻找的北大文史作为冷启动数据

Deepseek 作为开源模型，美国等大厂抵制最大的原因之一就是数据隐私，反过来中国应用开发者面对海外开源 LLM 时，也面临隐私性和中文创作能力问题。在创作领域，估计会有一大批个人创作者利用长推理的中文语言直接为日常工作/内容消费提供素材，国内 AI Fantasy、AI 网文、AI 短剧可能会迎来爆发式增长

- 开源/个人开发者平台 - Github、ProductHunt、Hugging Face

LCM 的兴起，来源于 R1 推理已经达到语义层级的思考，Next token prediction 可能不再受用，token 只是所有表征中的一种表征，用 token 的原因是工程上容易实现，然而人不用 token 方式思考，因为 token 是离散符号表达，无法代表连续且复杂的思维。近期大家讨论很多 Meta 提出的无需 tokenizer 的 BLT 架构与大型概念模型 LCM，直接在语义空间层面推理与生成，模型架构层可能出现变化

特征	传统LLM	LCM
抽象层级 (Level of Abstraction)	基于 Token层面工作	基于概念层面工作
输入表征 (Input Representation)	处理为单个 Token（子词）	处理为句子嵌入
输出生成 (Output Generation)	按单词逐字生成文本	按句子逐句生成文本
语言和模态支持 (Language and Modality Support)	通常针对特定语言进行训练	设计用于处理多种语言和模态
训练目标 (Training Objective)	最小化 Token 预测误差	最小化概念预测误差
推理与规划 (Reasoning And Planning)	隐式学习层次化推理	显式层次化推理
零试泛化能力 (Zero-Shot Generalizatin)	表现较弱	表现出色
长上下文处理效率 (Efficiency with Long Contexts)	效率较低	更高效
局限 (Limitation)	在深层次语义理解和多语言泛化方面存在局限。	训练存在模态竞争风险，计算效率较低；概念粒度不稳定；缺少专用的向量空间

(LLM vs LCM)

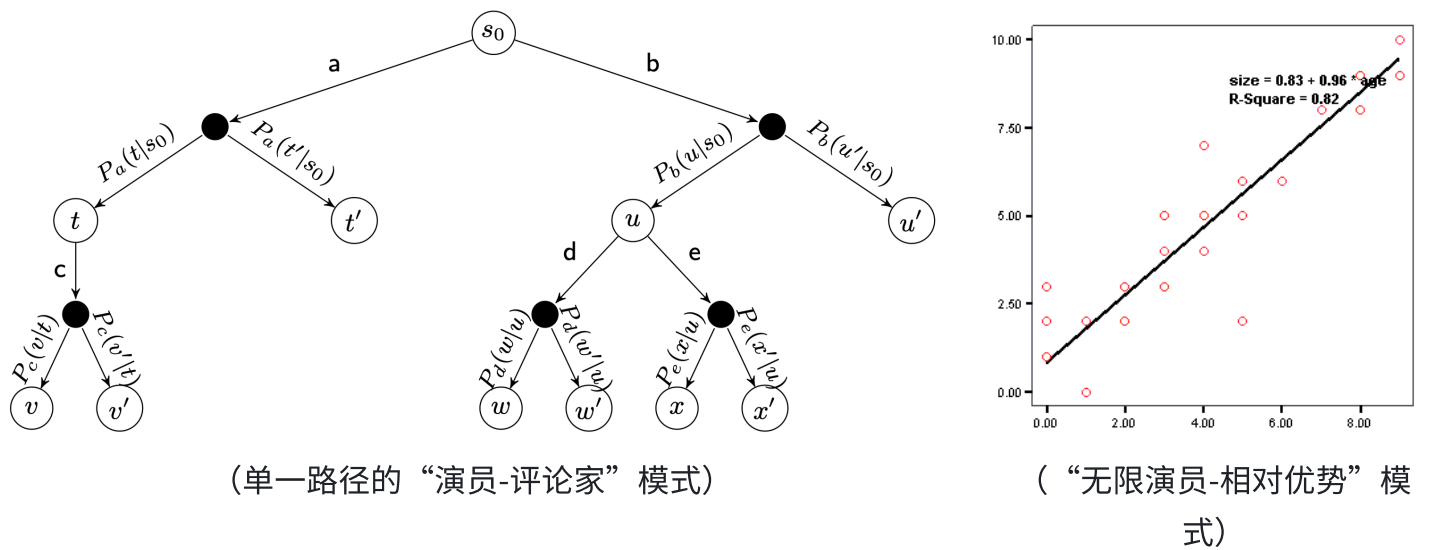
端侧，R1-1.5B 小模型非常适合手机部署，低价、数据隐私带来端侧模型的部署增加，华为开始与 Deepseek 合作

后训练的投入，R1 算法层最大的创新来自后训练中为行业带来的 RLHF 新方向，例如 AI for science 的深度推理在开源 R1 API 中能够有更多直接调用并测试可行性的机会。目前 R1 推理也有缺点，提的比较多的是过于“三心二意”，R1 因为会探索多条路径，频繁切换思路会带来额外 token 消耗，且仍然找不到正确答案，是否有测试错误答案，以及搭建高质量后训练奖励数据集的能力，会成为评判团队的重要指标之一

2.7 补充

R1的强化学习与传统强化学习有什么区别？

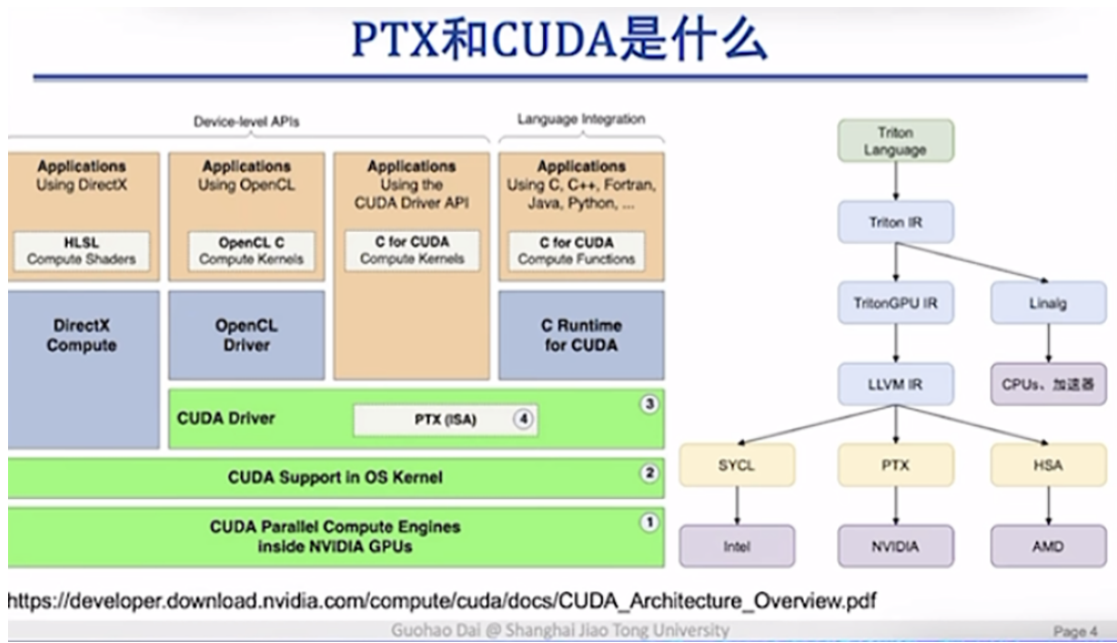
群体答案，相对优势，传统强化学习 PPO 类似“演员-评论家”，经严格训练的策略网络（演员）作出表演，价值网络（评论家）根据动作的期待回报打分。而 R1 采用群体相对策略优化 GRPO，统一输入生成多个候选回答，不再依赖独立训练的评论家，而是以这组演员的平均表演水平（统计信息，如平均奖励+标准差）作为标准，计算每个演员的相对优势，直观的反映同一问题下哪个回答更优，还减少了内存和计算开销



R1 利用 PTX 绕过了CUDA？

R1 并没有绕过 CUDA，而是采用了更底层的 CUDA 驱动层进行调用，R1 所用的 PTX 语言是定制并行线程指令，拥有自动调整通信块大小和缓存减少的特点，PTX 是 CUDA 驱动层的一个组件，由于处于

CUDA 编程接口之下，能够更方便的与底层硬件交互实现调用，主动优化互联，控制更多细节，包括通信优化、内存优化等底层优化，还能够兼顾 MLA 等算法协同优化，整体可优化空间变得更大



(PTX 仍然是 CUDA 生态中的驱动层)

Qwen-2.5 Max 超越了 R1?

Qwen-2.5 Max 学习了 R1 的强化学习思路 GRPO，在全球重要榜单 Chatbot Arena 上数学、编程能力超过 R1、o1，核心在于其后训练的一部分加入 GRPO，同样能够生成符合伦理和需求的响应。此外，Qwen-2.5 Max 更多是算法工程能力的加强，例如预训练上下文长度扩大，加入 RoPE 基频（生成不同频率位置编码，更好正弦/余弦频率定位，进行 token 相对位置捕捉），以及动态路由机制，根据历史知识激活相应专家等，对比可看出 Deepseek 对 infra 层的优化理解是独一无二的

但 Qwen-2.5 Max 不算是开源推理模型，他闭源，并没有公布模型权重和推理代码，同时也不像 R1 和 o1 一样能够看到思考过程，Qwen-2.5 Max 仍是一个通用模型

o1 对比 R1 做错了什么？

不开源，把o1深度思考过程隐藏起来，收费非常高，不能在全球让尽可能多的人去普惠、感受深度思考带来的震撼。R1 对标 o1，相当于 Llama 对标 ChatGPT，复现 o1 的困难在于训练数据，R1 的纯 RL 生成出来的结果让人产生非常多遐想，在这一点创新上行业会思考，未来是不是可以通过大规模强化学习在推理上超过 o1

年轻人用 R1 做什么小生意？

春节赚5k-1w，旅游攻略、短视频文案（脚本、解说字幕）、情感咨询，哪吒电影二创（敖丙+哪吒同人文） ...