

DeepSeek-R1 应用

2025.2.23

CONTENTS 目录

第一章

DeepSeek-R1 的特点

第二章

DeepSeek-R1 的后续

第三章

For AI for Science ?

第四章

总结



Part

1

DeepSeek-R1的特点



训练过程

通用大模型的基本训练过程：

预训练 (PreTrain)

后训练 (PostTrain)

推理 (Inference)

“Pre-training as we know it will unquestionably end”

----- Ilya Sutskever

背景：

o1 让模型具备了超强的推理能力，具体实现细节没有公开，只透露了Test-time Scaling这种方法（推理阶段）。

DeepSeek-R1的贡献和训练过程 (Jan 22)：

通过后训练阶段的优化，复现了o1的性能和超强推理能力，并且开源。

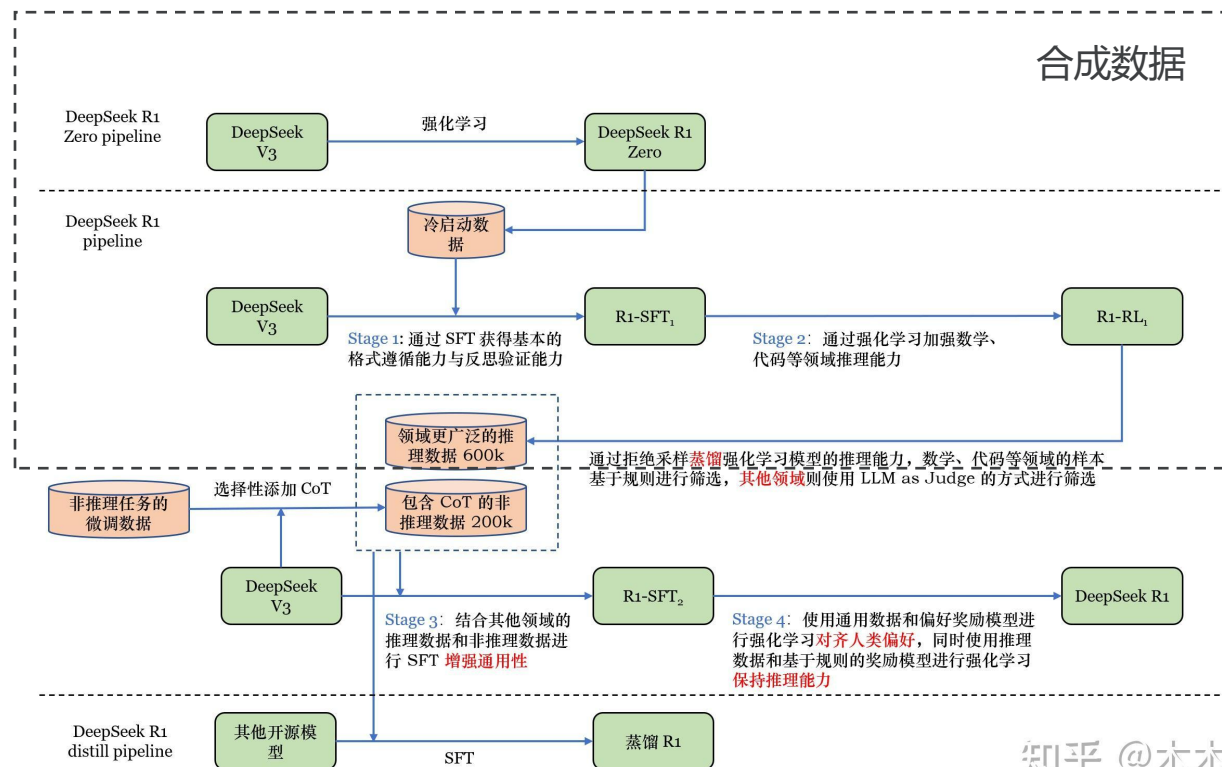
训练基础：DeepSeek-V3，预训练模型

训练阶段1：纯强化学习训练（能够思考，但是可读性差）

训练阶段2：冷启动（少量Long-CoT微调），继续RL

训练阶段3：生成数据。和部分真实数据混合，SFT

训练阶段4：继续RL，基于规则的RL和RLHF。





核心特点总结

特点一：

纯强化学习（RL）驱动的推理能力（Aha Moment）：DeepSeek R1（尤其是R1-Zero版本）完全跳过了传统的监督微调（SFT），**通过强化学习直接从基础模型中发展出推理能力**。它采用“试错学习”模式，结合可验证的奖励机制（例如正确性奖励、格式奖励、语言一致性奖励），让模型自主探索推理路径。这种方法**减少了对大规模标注数据集的依赖**，同时提升了模型在数学、编码等可验证领域的性能。

特点二：

多阶段训练pipeline：R1的训练分为多个阶段：从冷启动阶段（使用少量高质量数据初始化）到多轮强化学习优化（包括数学推理精炼和人类偏好对齐）。这种结构化的训练流程确保了推理能力的逐步增强和输出的可读性。

特点三：

合成的高质量数据：R1训练的前两个阶段都是为了合成高质量的Long-CoT推理数据（大幅降低了对标注的依赖），这里的**推理链**非常重要，是模型实现推理能力的一个关键。

特点四：

混合专家架构（MoE）与高效推理：R1继承了V3的MoE架构，总参数671B，但每token仅激活37B参数。这种架构通过稀疏激活大幅减少了计算需求，同时结合多头潜在注意力（Multihead Latent Attention）和多词生成技术提升推理效率。

特点五：

高效的模型蒸馏（Distillation）：DeepSeek R1不仅推出了旗舰模型，还通过蒸馏技术生成了多个较小模型（如1.5B到70B参数的版本），这些模型在保持推理能力的同时大幅降低了计算需求。例如，其32B参数版本在AIME 2024中达到72.6%准确率，超越了许多更大的开源模型。



Self-evolution 和 Aha Moment

Self-evolution

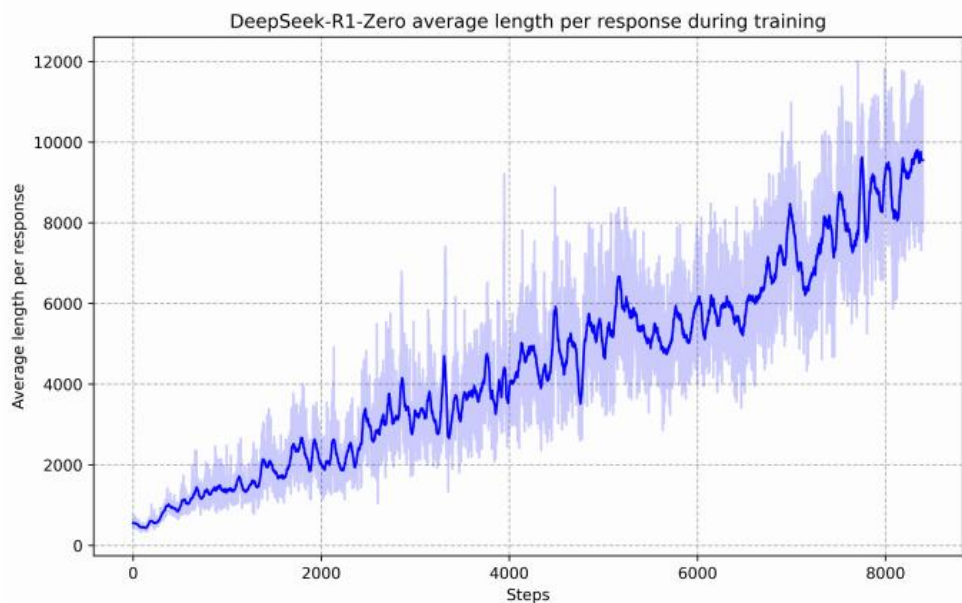


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

随着训练轮数增加，模型回答长度逐渐变长，并且出现了类似于反思的过程。这些行为不是明确编程的，而是模型与强化学习环境的相互作用的结果。

Aha Moment

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: ...

...

Table 3 | An interesting “aha moment” of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

重新评估其初始方法来学习为一个问题分配更多的思考时间。这种行为不仅证明了该模型不断增长的推理能力，也是一个迷人的例子，说明强化学习如何导致意想不到和复杂的结果。



Part **2**

DeepSeek-R1的后续

在顿悟时刻，模型会习得诸如自我反思这类突发技能，这有助于它进行上下文搜索来解决复杂的推理问题。在R1-Zero发布后的短短几天内，就有几个项目独立地在较小规模（例如，10亿到70亿参数规模）上“复现”了类似R1-Zero的训练，并且都观察到了顿悟时刻，通常以**模型响应长度突然增加**来衡量。我们遵循他们的设置来仔细研究类似R1 - 零的训练过程，并分享以下发现：

- 在类似R1-Zero的训练中可能不存在顿悟时刻。相反，我们发现顿悟时刻（例如自我反思模式）出现在起始轮次（epoch 0），也就是基础模型阶段。
- 我们发现基础模型的响应中存在**表面自我反思**（SSR），在这种情况下，自我反思不一定能得出正确的最终答案。
- 我们更深入地研究了通过强化学习进行的类似R1-Zero的训练，发现**响应长度增加这一现象并非是因为自我反思的出现，而是强化学习优化精心设计的基于规则的奖励函数的结果。**

总结：

- RL是将原本表面的自我反思转化为有效的自我反思，以最大化预期奖励，从而提高推理能力。
- 长度和自我反思可能不相关。
- 总的来说，可以概括成两句话：Base模型也可能Aha，但不否认RL不能Aha；RL能将Base的表面自我反思转化为有效自我反思，只是并不一定长度就一定增加。我觉得这个结论是Make sense的，Base模型只是具有能力但没有被激活，RL才激活了能力。

- 在类似R1-Zero的训练中可能不存在顿悟时刻。相反，我们发现顿悟时刻（例如自我反思模式）出现在起始轮次（epoch 0），也就是基础模型阶段。

Model	Best template	Observed self-reflection keywords in model responses
Qwen2.5-Math-1.5B, Qwen2.5-Math-8B, Qwen2.5-7B	Template 2	rethink, recheck, try again, let's correct it and verify the steps again.
Microsoft-Rho-Math-7B	Template 2	recheck, re-evaluate, check again, try again.
DeepSeek-Math-7B-Base	Template 1	let's try again, let's think again.
Llama-3.1-8B	Template 1	None

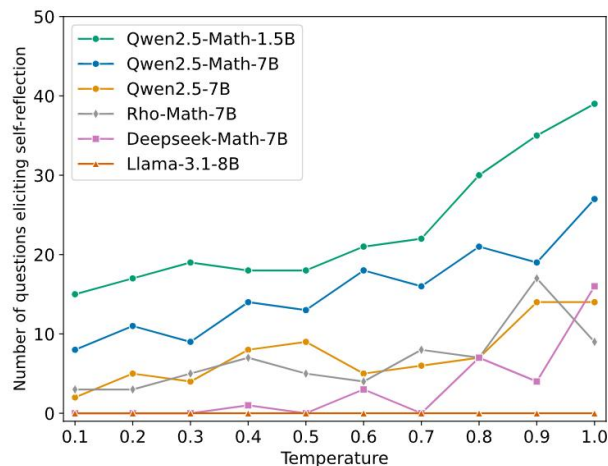


Fig 1a. Number of questions out of 500 MATH questions eliciting self-reflection behaviors in different base models.

在不同基础模型中引发自我反思行为的问题数量。结果表明，在不同温度下均可观察到自我反思现象，并且呈现出温度越高，在第0轮（epoch 0）出现“顿悟时刻”（Aha moment）的频率越高的趋势。

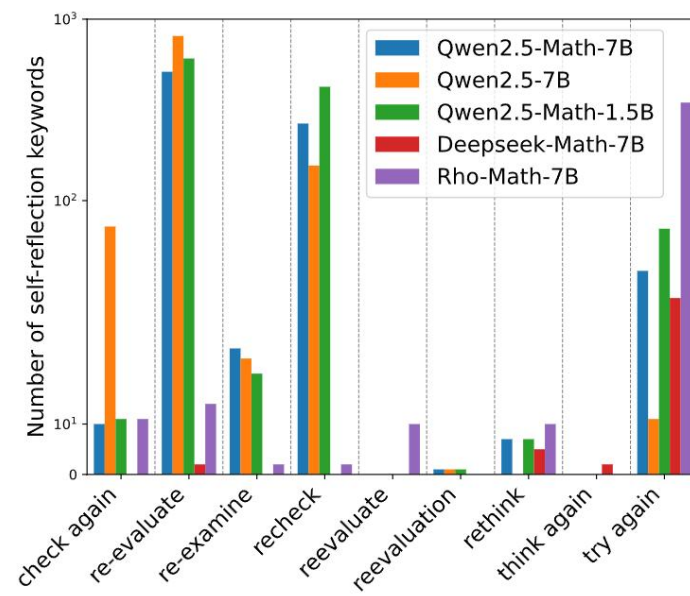
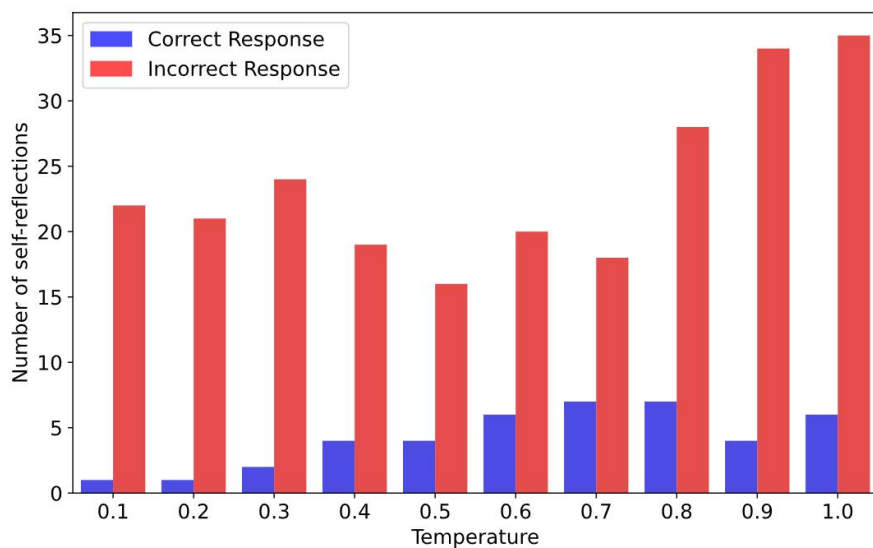


Fig 1b. Number of keyword occurrences out of 40,000 responses (500 questions \times 8 responses per question \times 10 temperatures). y is in log scale.

- 我们发现基础模型的响应中存在**表面自我反思**（SSR），在这种情况下，自我反思不一定能得出正确的最终答案。

并非所有来自基础模型自我反思都是能带来改进解决方案的有效反思。为便于讨论，我们将它们定义如下，称之为**表面性的自我反思**。

表面自我反思（SSR）是指模型回应中缺乏建设性修订或改进的重新评估模式。与没有自我反思的回应相比，表面自我反思不一定能得出更好的答案。



正确与错误回答中自我反思的数量。蓝色柱形代表正确回答中自我反思关键词出现的总次数，红色柱形代表错误回答中自我反思关键词出现的总次数。

在不同的采样温度下，大多数自我反思（以其频率衡量）并未导致正确答案，这表明基础模型容易产生表面的自我反思。

Fig 4. Number of self-reflections in correct and incorrect responses. The blue bars represent the total occurrences of self-reflection keywords in correct responses, while the red bars represent the total occurrences in incorrect responses.



复现R1的计算量比较大， $\geq 32K$ context， ~ 8000 steps，即使对1.5B模型也需要70,000小时A100。为了解决这个问题，本文使用一个蒸馏模型，并引入迭代式的长度增加方案。计算资源降到3800小时A100。

DeepScaleR-1.5B-Preview

Model	AIME 2024	MATH 500	AMC 2023	Minerva Math	Olympiad Bench	Avg.
DeepScaleR-1.5B-Preview	43.1	87.8	73.6	30.2	50.0	57.0
DeepSeek-R1-Distill-Qwen-1.5B	28.8	82.8	62.9	26.5	43.3	48.9
O1-Preview	40.0	81.4	-	-	-	-

直接从[Deepseek-R1-Distilled-Qwen-1.5B](#)强化微调（用**高质量40K Math数据**）。得到的模型在AIME2024和MATH500上超过了[o1-preview](#)。

好消息：**1.5B蒸馏模型上进行的强化学习微调就可以实现Scaling!**

我们朝着揭示利用强化学习（RL）将小模型转变为强大推理模型的方法迈出了一步。我们介绍了DeepScaleR - 1.5B - 预览版，该模型使用4万个高质量的数学问题进行训练，消耗3800个A100 GPU小时（花费4500美元, ），在多个竞赛级别的数学基准测试中优于OpenAI的o1 - 预览版



两步训练：（先短后长，错误响应的长度是正确响应的3倍。这表明较长的响应通常会导致不正确的结果，直接用长上下文窗口进行训练可能效率低下）

- 先在8k上下文训练，平均训练奖励从 46% 增加到 58%，而平均响应长度从 5,500 下降到 3,500，AIME2024 Pass@1 Acc 33.9%。1000步后，响应长度再次开始增加，响应裁剪率从 4.2% 上升到 6.5%，表明更多的响应在上下文限制处被截断。说明此时模型试图通过“思考更久”（即生成更多Token）来提高训练奖励。
- 在第 1040 步（响应长度开始呈上升趋势）处使用 16K 上下文窗口重新启动训练。额外 500 步后，平均响应长度从 3500 增加到 5500，平均训练奖励稳定到62.5%，Acc达到38%，响应裁剪率到2%。性能开始趋于稳定。

	Step1	Step2	Step2-add
ContextLength	8k	16k	24k
ResponseLength	5500→3500	3500→5500	
ClipRatio	4.2%→6.5%	2%	
Reward	46%→58%	62.5%	
AIME Acc	33.9%	38%	43.1%

总的来说，结论就是：

- RL 缩放也可以表现在小型模型中。单独的 SFT 和 RL 都不够；相反，通过将高质量的 SFT 蒸馏与RL缩放相结合，可以真正释放 LLM 的推理潜力。AIME 准确率从 28.9% 提高到 43.1%。
- 迭代扩展长度可实现更有效的长度缩放。简单来说，就是先训短的简单的，然后提升难度和长度限制，这样会比直接训练更加高效。



LIMO: Less is More for Reasoning

传统观点认为，复杂的推理任务需要大量的训练数据（通常 $>100,000$ 个示例），但我们展示了一个惊人的现象：**复杂的数学推理能力可以通过少得惊人的示例有效地被激发出来**。这一发现不仅挑战了大量数据需求的假设，也挑战了监督微调主要导致记忆而非泛化的普遍看法。

仅使用817个精心策划的训练样本，我们提出的LIMO模型在极具挑战性的AIME基准测试中达到了57.1%的准确率，在MATH测试中达到了94.8%的准确率，将之前强大的基于SFT（监督微调）的模型在AIME上的性能从6.5%提高到**57.1%**，在MATH上的性能从59.2%提高到94.8%，同时仅使用了之前方法所需训练数据的1%。

最值得注意的是，LIMO展现出了卓越的分布外泛化能力，在10个不同的基准测试中实现了40.5%的绝对提升，超过了使用100倍更多数据训练的模型，直接挑战了SFT本质上导致记忆而非泛化的普遍观念。

综合这些开创性的结果，我们提出了“**少即是多推理假说（LIMO假说）**”：**在预训练期间领域知识已被全面编码的基础模型中，复杂的推理能力可以通过最少但精心编排的认知过程演示来出现。**

该假说认为，**复杂推理的激发阈值**并非由目标推理任务的复杂性所固有地限制，而是由两个关键因素从根本上决定的：

- (1) 模型在预训练期间编码的知识基础的完整性；**(强大的预训练模型)**
- (2) 训练后示例的有效性，这些示例充当“认知模板”，向模型展示如何有效地利用其现有的知识库来解决复杂的推理任务。



LIMO: Less is More for Reasoning

我们将少即是多推理（LIMO）假设正式阐述如下：

在**预训练期间已全面编码领域知识的基础模型**中，通过最小化但精心编排的认知过程演示，可以产生复杂的推理能力。该假设基于两个基本前提：

- （一）模型参数空间内存在先验知识的潜在存在；
- （二）推理链的质量，即将复杂问题精确分解为详细的、逻辑性的步骤，使认知过程明确且可追溯。

为了验证这一假设，我们提出了一种**构建高质量、最小化数据集**的系统方法，该方法可有效激发模型固有的推理能力。

高质量数据整理：

3.3.1 问题选择

我们假设高质量的问题（ q 属于 Q ）自然会引发深入的推理过程。我们的选择标准包括以下几点：

- **难度水平**：我们优先考虑具有挑战性的问题，这些问题能够促进复杂的推理链、多样化的思维过程和知识整合，使大型语言模型（LLMs）能够有效地利用预训练知识进行高质量的推理。
- **普适性**：与模型的训练分布偏离较大的问题能够更好地挑战其固定的思维模式，鼓励探索新的推理方法，从而扩大其推理搜索空间。
- **知识多样性**：所选问题应涵盖各种数学领域和概念，要求模型在解决问题时整合和连接不同的知识。



LIMO: Less is More for Reasoning

高质量数据整理：

3.3.1 问题选择

我们假设高质量的问题（ q 属于 Q ）自然会引发深入的推理过程。我们的选择标准包括以下几点：

- **难度水平：** 我们优先考虑具有挑战性的问题，这些问题能够促进复杂的推理链、多样化的思维过程和知识整合，使大型语言模型（LLMs）能够有效地利用预训练知识进行高质量的推理。
- **普适性：** 与模型的训练分布偏离较大的问题能够更好地挑战其固定的思维模式，鼓励探索新的推理方法，从而扩大其推理搜索空间。
- **知识多样性：** 所选问题应涵盖各种数学领域和概念，要求模型在解决问题时整合和连接不同的知识。

具体做法：

首先，从数千万个题目组成的初始题库开始，我们使用Qwen2.5 - Math - 7B - Instruct（杨等人，2024年）应用了一个基础难度筛选器，排除了该模型能够在数次尝试内正确解答的题目。这个过程有助于确定一个初步的难度阈值。

随后，我们对剩余的题目使用包括R1、DeepSeek-R1-Distill-Qwen32B（郭等人，2025年）以及黄等人（2024年）的模型在内的最先进的推理模型进行了更严格的评估，仅保留那些即使是最强大的模型经过多次采样迭代后成功率低于某个阈值的题目。

最后，为了保持语料库的多样性，我们采用了策略性采样技术，在数学领域和复杂程度方面平衡代表性，同时避免概念性重复。这一细致的筛选过程最终从数千万个候选题目中精心挑选出了817个题目，所选题目整体满足我们严格的质量标准，并且涵盖了丰富的数学推理挑战类型。



LIMO: Less is More for Reasoning

高质量数据整理：

3.3.2 推理链构建

除了高质量的问题之外，解决方案的质量在大语言模型的训练阶段也起着关键作用。为了整理出高质量的解决方案，我们采用了一种全面的筛选策略。

我们首先收集了有官方解决方案的问题（如有），并辅以人类专家和人工智能专家撰写的解决方案。此外，我们还利用了最先进的推理模型，包括DeepSeek-R1、DeepSeek-R1-Distill-Qwen32B（郭等人，2025年）和Qwen2.5-32b-Instruct，来生成多样化的解决方案思路。

随后，所有作者通过协作审查对这些筛选后的解决方案进行了全面分析。通过仔细观察和系统评估，我们确定了几个区分高质量推理链的关键特征。

- **最优结构组织：**该解决方案呈现出清晰且组织良好的结构格式，在步骤分解方面具有自适应的粒度。特别是，在关键推理节点分配更多标记并进行详细阐述，而对于简单步骤则保持简洁表达。这种针对步骤粒度的自适应方法确保复杂转换得到适当关注，同时避免在较简单的推导中出现不必要的冗长表述。
- **有效的认知支架：**高质量的解决方案通过精心构建的解释逐步构建理解，从而提供策略性的教育支持。这包括概念的渐进式引入、在关键点清晰阐述关键见解以及巧妙弥合理念差距，使复杂的推理过程更易于理解和掌握。
- **严格的验证：**高质量的解决方案在整个推理过程中包含极为频繁的验证步骤。这包括验证中间结果、交叉检查假设以及确认每个推导的逻辑一致性，从而确保最终答案的可靠性。



s1: Simple test-time scaling

测试时缩放是一种用于语言建模的有前景的新方法，它利用额外的测试时计算资源来提升性能。近期，OpenAI的o1模型展现了这种能力，但没有公开其方法，这导致了許多复制尝试。我们寻求实现测试时缩放和强大推理性能的最简方法。首先，我们精心整理了一个包含1000个问题且搭配推理轨迹的小数据集s1K，这些问题依据我们在消融实验中验证过的三个标准（难度、多样性和质量）进行挑选。

其次，我们开发了预算强制机制来控制测试时计算资源，具体做法是强制终止模型的思考过程，或者当模型试图结束时通过在模型的生成内容后多次添加“等待”来延长该过程。这会使模型复查自己的答案，常常修正错误的推理步骤。

在对Qwen2.5 - 32B - Instruct语言模型使用s1K数据进行监督微调并为其配备预算强制机制之后，我们的s1 - 32B模型在竞赛数学问题（如MATH和AIME24）上比o1 - preview模型的表现提升了多达27%。此外，通过预算强制机制对s1 - 32B进行扩展能够在其无测试时干预的性能基础上进行外推：在AIME24上的表现从50%提升到57%。

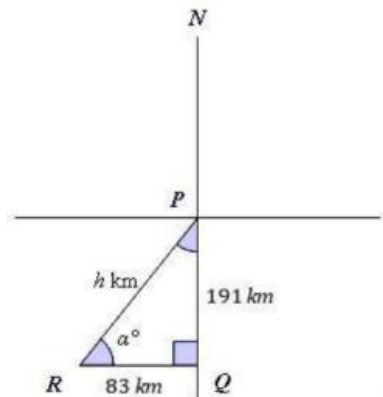


Large Vision-Language Model?

Multimodal Open R1

<https://github.com/EvolvingLMMs-Lab/open-r1-multimodal>

We conducted a speed-run on to investigate R1's paradigm in multimodal models after observing growing interest in R1 and studying the elegant implementation of the GRPO algorithm in `open-r1` and `tr1`.



Problem

Based on the image, what is the ratio of the leg adjacent to the right angle (83 km) to the hypotenuse (h km)? Choices: A. 1:2 B. 1:3 C. 1:4 D. 2:3



Completion #1

<think>
Finding the ratio of the segment adjacent...Let's denote the hypotenuse as h km and the segment adjacent to the right angle as 83 km...</think>
<answer> B. 1:3 </answer>

Reward #1

1

Completion #8

We can use the formula $\sin(A) = \text{opposite} / \text{hypotenuse}$. Here, the opposite side is $5x$ and the hypotenuse is $5x$. So, $\sin(A) = 5x / 5x = 1$.

Reward #8

0

Solution

<think>Let me think about this. The image shows a right triangle with one leg measuring 83 km and the hypotenuse measuring 191 km. To find the ratio of the leg adjacent to the right angle to the hypotenuse, I need to divide the length of the leg by the length of the hypotenuse. So, the ratio is $83 \text{ km} / 191 \text{ km}$. Simplifying this fraction, I find that it approximately equals $1/3$. Therefore, the ratio is 1:3.</think>
<answer>B</answer>

The training commands below are configured for a node of 8 x H100s (80GB). For different hardware and topologies, you may need to tune the batch size and number of gradient accumulation steps.

Question

Model

GRPO Process



Qwen2.5-VL (Feb 20)

我们推出Qwen2.5 - VL，这是Qwen视觉 - 语言系列的最新旗舰模型，其在基础能力和创新功能方面均有显著提升。

从技术层面来看，我们的贡献体现在四个方面：

- (1) 我们在视觉编码器中实现了窗口注意力机制，以优化推理效率；
- (2) 我们引入了动态帧率采样，将动态分辨率扩展到时间维度，从而实现针对不同采样率的全面视频理解；
- (3) 我们通过将时间域中的MRoPE（相对位置编码）与绝对时间对齐来对其进行升级，进而促进更复杂的时间序列学习；
- (4) 我们在预训练和有监督微调的数据整理方面付出了巨大努力，将预训练语料库的规模从1.2万亿个词元进一步扩大到4.1万亿个词元。

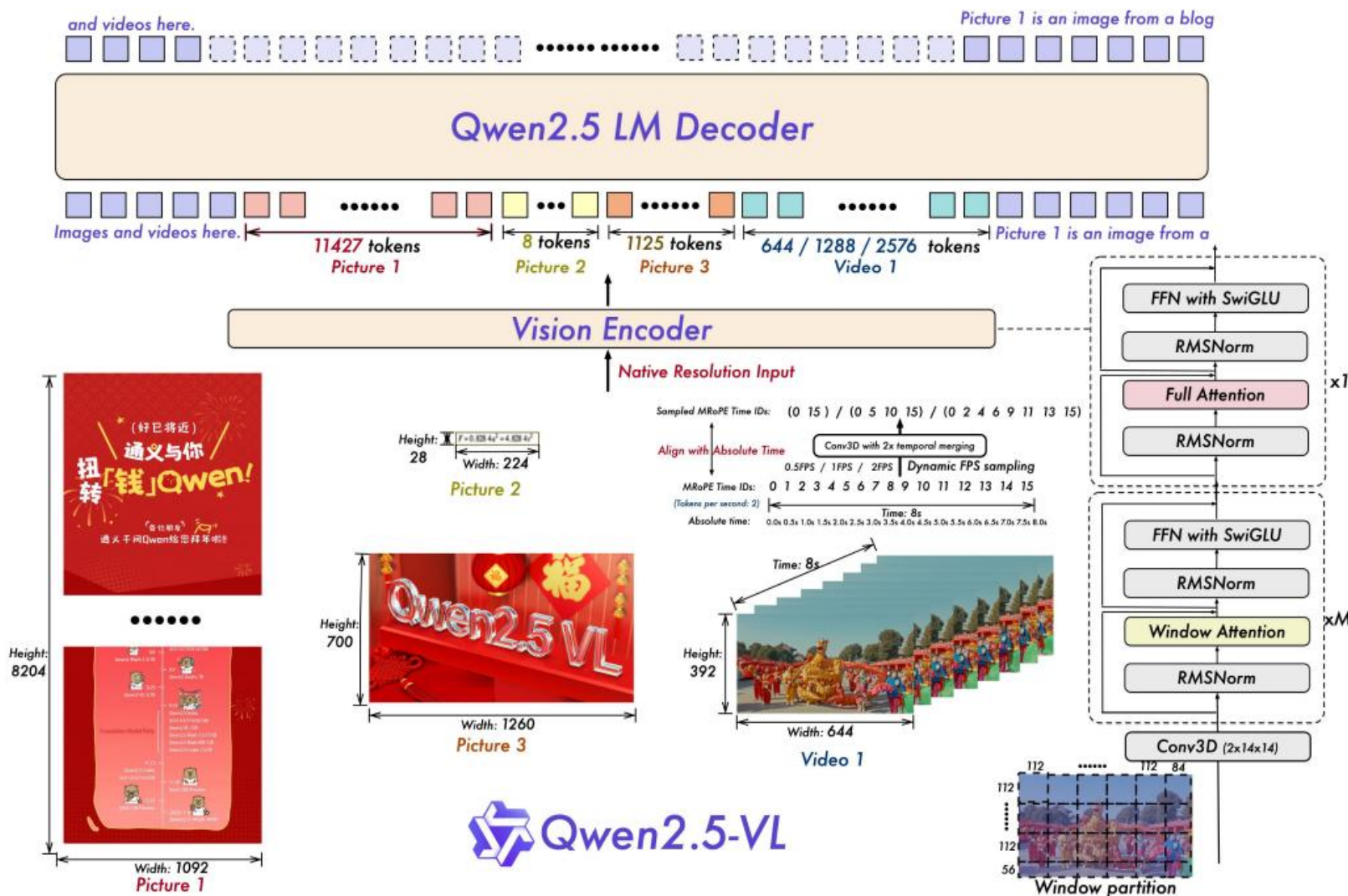
特点：

- (1) 强大的文档解析能力：文心一言2.5 - 视觉版（Qwen2.5 - VL）将文本识别升级为全文档解析，在处理多场景、多语言以及各种内置（手写、表格、图表、化学公式和乐谱）文档方面表现卓越。
- (2) 跨格式的精确对象定位：文心一言2.5 - 视觉版（Qwen2.5 - VL）在检测、指向和计数对象方面实现了更高的准确性，可适应绝对坐标和JSON格式以实现高级空间推理。
- (3) 超长视频理解与细粒度视频定位：我们的模型将原生动态分辨率扩展到时间维度，增强了理解长达数小时的视频的能力，同时能够提取以秒计的事件片段。
- (4) 增强的计算机和移动设备上的智能体功能：利用先进的定位、推理和决策能力，为智能手机和计算机上的模型提升更强大的智能体功能。

该模型无需针对特定任务进行微调就能在各个领域实现强大的泛化能力。Qwen2.5 - VL有三种尺寸可供选择，可满足从边缘人工智能到高性能计算的多样化用例需求。**旗舰型号Qwen2.5 - VL - 72B模型可与GPT - 4o和Claude 3.5 Sonnet等最先进的模型相媲美**，尤其在文档和图表理解方面表现卓越。较小的Qwen2.5 - VL - 7B和Qwen2.5 - VL - 3B模型比同类竞品表现更优，即使在资源受限的环境下也能提供强大的功能。此外，Qwen2.5 - VL保持了强大的语言性能，保留了Qwen2.5 LLM的核心语言能力。



Qwen2.5-VL (Feb 20)



Qwen2.5 - VL框架展示了视觉编码器和语言模型解码器的集成，以处理包括图像和视频在内的多模态输入。

视觉编码器旨在以其原始分辨率处理输入，并支持动态每秒帧数（FPS）采样。不同大小的图像和具有不同FPS率的视频帧被动态映射到不同长度的标记序列。

值得注意的是，MRoPE沿时间维度将时间标识与绝对时间对齐，使模型能够更好地理解时间动态，例如事件的节奏和精确的时刻定位。处理后的视觉数据随后被送入Qwen2.5语言模型（LM）解码器。我们重新设计了视觉Transformer（ViT）架构，纳入了诸如带有SwiGLU激活函数的前馈网络（FFN）、用于归一化的RMSNorm以及基于窗口的注意力机制等先进组件，以提高性能和效率。



Qwen2.5-VL (Feb 20)

Configuration	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen2.5-VL-72B
Vision Transformer (ViT)			
Hidden Size	1280	1280	1280
# Layers	32	32	32
# Num Heads	16	16	16
Intermediate Size	3456	3456	3456
Patch Size	14	14	14
Window Size	112	112	112
Full Attention Block Indexes	{7, 15, 23, 31}	{7, 15, 23, 31}	{7, 15, 23, 31}
Vision-Language Merger			
In Channel	1280	1280	1280
Out Channel	2048	3584	8192
Large Language Model (LLM)			
Hidden Size	2048	3,584	8192
# Layers	36	28	80
# KV Heads	2	4	8
Head Size	128	128	128
Intermediate Size	4864	18944	29568
Embedding Tying	✓	✗	✗
Vocabulary Size	151646	151646	151646
# Trained Tokens	4.1T	4.1T	4.1T

Table 1: Configuration of Qwen2.5-VL.

Stages	Visual Pre-Training	Multimodal Pre-Training	Long-Context Pre-Training
Data	Image Caption Knowledge OCR	+ Pure text Interleaved Data VQA, Video Grounding, Agent	+ Long Video Long Agent Long Document
Tokens	1.5T	2T	0.6T
Sequence length	8192	8192	32768
Training	ViT	ViT & LLM	ViT & LLM

Table 2: Training data volume and composition across different stages.



Qwen2.5-VL (Feb 20)

Configuration	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen2.5-VL-72B
Vision Transformer (ViT)			
Hidden Size	1280	1280	1280
# Layers	32	32	32
# Num Heads	16	16	16
Intermediate Size	3456	3456	3456
Patch Size	14	14	14
Window Size	112	112	112
Full Attention Block Indexes	{7, 15, 23, 31}	{7, 15, 23, 31}	{7, 15, 23, 31}
Vision-Language Merger			
In Channel	1280	1280	1280
Out Channel	2048	3584	8192
Large Language Model (LLM)			
Hidden Size	2048	3,584	8192
# Layers	36	28	80
# KV Heads	2	4	8
Head Size	128	128	128
Intermediate Size	4864	18944	29568
Embedding Tying	✓	✗	✗
Vocabulary Size	151646	151646	151646
# Trained Tokens	4.1T	4.1T	4.1T

Table 1: Configuration of Qwen2.5-VL.

R1-V: Reinforcing Super Generalization Ability in Vision Language Models with Less Than \$3

贡献:

- 我们首先揭示了具有可验证奖励的强化学习 (RLVR) 在视觉语言模型的有效性和分布外 (OOD) 鲁棒性方面均优于思维链监督微调 (CoT - SFT) 。
- 在我们的实验中, 我们激励视觉语言模型学习可泛化的视觉计数能力, 而非过度拟合训练集。
- 在仅100个训练步骤内, 20亿参数的模型在OOD测试中的表现优于720亿参数的模型。
- 此次训练是在8个A100 GPU上进行的, 耗时30分钟, 花费2.62美元。

问题:

具有可验证奖励的强化学习 (RLVR) 在分布外 (OOD) 视觉推理任务 (例如, 复杂计数/问答) 中的表现优于监督微调 (SFT), 而SFT在诸如几何图形之类的领域内场景中表现出色。

在SFT中强制思维链 (Chain - of - Thought) 推理往往会损害小型视觉语言模型 (VLMs) 的性能。RLVR的GRPO方法能够在无人类指导的情况下实现自适应推理 (无思维链/长思维链), 从而实现强大的泛化能力 (在SuperClevr数据集上比SFT高出68.7%对比19.4%) 。

- 关键的开放性问题:
 - 为什么RL/SFT在不同领域中各有优势? 如何将它们结合起来?
 - 模型规模如何影响RL/SFT的性能差异?
 - 如何为开放式视觉/智能体任务设计奖励?

未来的工作旨在统一这些方法, 并将RL扩展到结构化程度较低的问题。



Part **3**

For AI for Science ?



MedS3 (Feb 17)

MedS³: Towards Medical Small Language Models with Self-Evolved Slow Thinking

Shuyang Jiang^{♣,◇}, Yusheng Liao^{♠,◇}, Zhe Chen^{♠,◇},
Ya Zhang^{♠,◇}, Yanfeng Wang^{♠,◇}, Yu Wang^{*,♠,◇}

[♠]Shanghai Jiao Tong University

[♣]Fudan University

[◇]Shanghai Artificial Intelligence Laboratory

shuyangjiang23@m.fudan.edu.cn

{liao20160907,chenzhe2018,ya_zhang,wangyanfeng622,yuwangsjtu}@sjtu.edu.cn

背景:

以往的医疗模型训练面临医疗语料匮乏的问题，通常有两种解决方案：

- (1) 在大规模人工收集筛选的医疗语料上进行预训练；
- (2) 在少量特定任务数据集上进行有监督微调。

然而，第一种方法消耗大量计算资源，但下游任务性能提升有限；第二种方法虽计算高效，但微调数据多为闭源模型生成的蒸馏数据或人工标注的短回复数据，限制了模型的优化空间和跨任务泛化能力。

核心贡献:

- MedS3 由策略模型 (Policy Model) 和过程奖励模型 (Process Reward Model; PRM) 组成，通过在 16 种不同数据集上的学习，包括医疗诊断、生物医学和知识性问答等。
- 仅使用 7465 条种子数据，结合细粒度的蒙特卡洛树搜索和规则验证的过程监督信号，MedS3 迭代优化策略模型和过程奖励模型。
- MedS3 在医疗知识问答、生物医学问答、长上下文问答和医疗诊断任务上的推理能力显著超越现有医疗大模型和通用域推理模型，成为首个在医疗诊断任务上实现长链推理「R1」的大语言模型框架。



MedS3 (Feb 17)

强化学习的一个简单背景：在强化学习的框架下重新看大推理模型，它有四个重要因素：

一是策略初始化 (Policy Initialization)，通过预训练、提示工程、监督微调，让模型具有初始的类人推理行为，比如问题理解、任务分解、验证修正错误等。

二是奖励设计 (Reward Design)，为RL提供奖励信号。传统方法分为两种，一是从环境直接获取奖励信号，二是从专家数据或者偏好数据学习奖励。o1应该是混合了多种奖励设计的方法。比如有groundtruth的环境，将结果监督 (ORM) 转换为过程监督 (PRM)。没有groundtruth的话，就用专家或者偏好数据来学习奖励。在大量领域上训练奖励模型，提升泛化性。

三是搜索 (Search)，找寻问题的最优解法。传统方法基本分为两大类，基于树的搜索和基于顺序修改的搜索。这两种可能对复现o1都有非常大的帮助。

四是学习 (Learning)，优化模型参数。基本上就是一个是用强学习的PolicyGradient，还有一个BehaviorCloning。这两种基本上可以用在两个阶段：Warmup阶段可以使用行为克隆方法，快速收敛；第二阶段再用强化学习来提升上限。

为了解决医疗模型的数据困境，MedS3 转向运行时缩放 (test-time scaling)，以**一种数据高效的后训练方法进行提升，从而突破数据集标注的约束，在平衡计算资源与性能之间的矛盾下，高效利用现有的医疗数据。**

MedS3 的核心在于其独特的**自我进化框架**。

研究者首先利用蒙特卡洛树搜索 (MCTS) 技术，基于基础策略模型生成可验证的推理链。在推理链的每一步，都会基于这一步的正确性赋予一个展开值，通过这些经过验证的轨迹来训练策略模型和过程奖励模型 (PRM)。

这种搜索对计算资源的依赖极小，通过策略模型演化得到的正负样本均可以作为 MedS3 的监督信号，**大大增加了数据利用率，并且按步采样也能提升模型的探索空间。**



MedS3 (Feb 17)

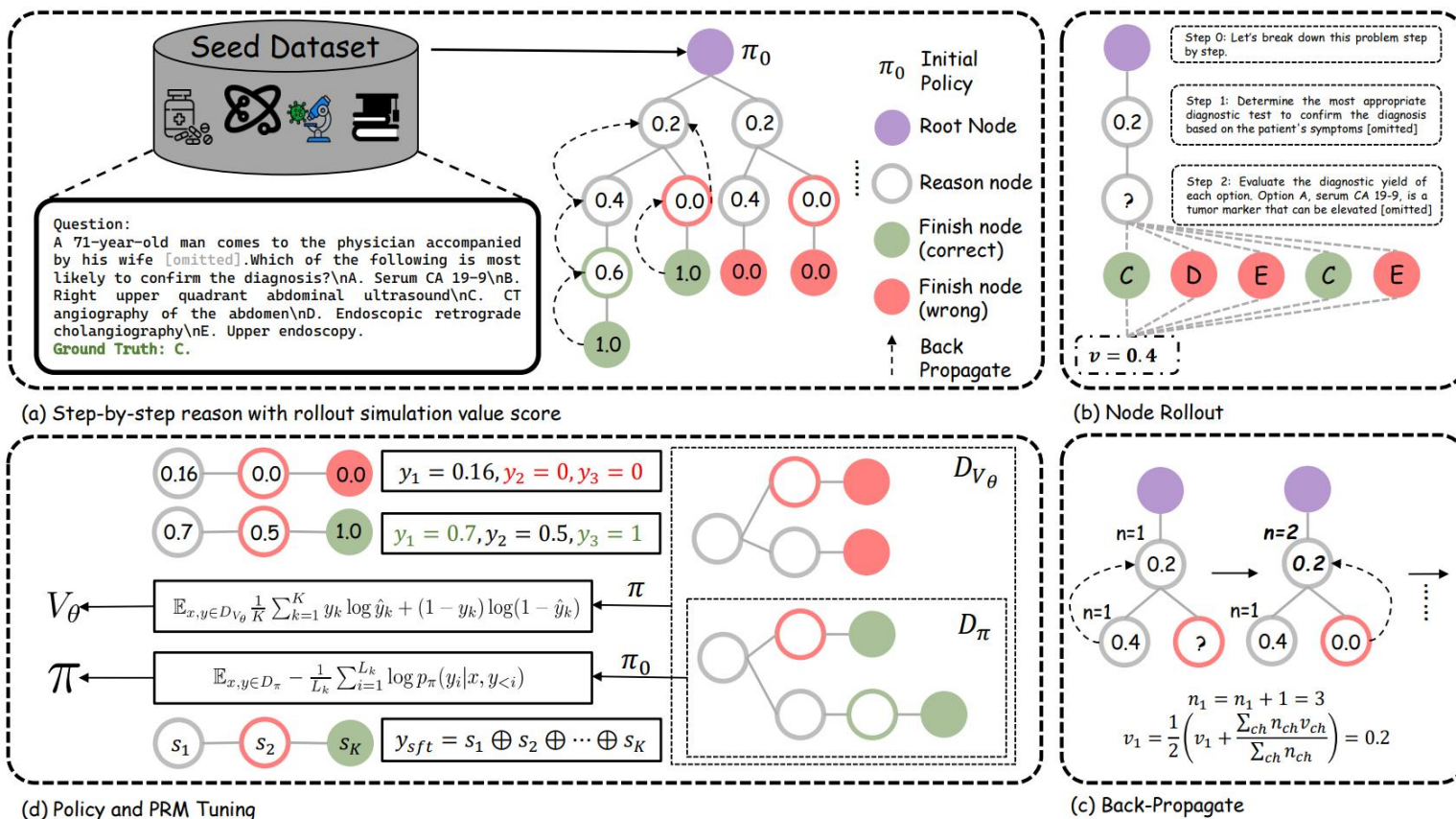


图1：MedS3框架构建的概览。
MedS3利用蒙特卡洛树搜索流程为种子数据集中的每个问题自动生成逐步推理路径（a）。

在此过程中，MedS3使用结果模拟来获取每个节点的展开值（b）；在获得子节点的展开值后，MedS3执行反向传播，以便从更深层精确地进行价值预测并回传到较浅的节点（c）。

在收集所有正确和错误的结束节点后，我们使用监督微调来优化策略模型 π ，采用正确的推理轨迹和逐步判别损失来获得过程奖励模型 V_θ （d）。



MedS3 (Feb 17)

Models	Knowledge QA		BioMedical QA			Long Context QA		NLI		Diagnosis QA		Avg.
	MedQA	MedMCQA	PubMedQA	BioASQ	MMLU	BioMRC	DDXPlus	PubHealth	HealthFact	DrugDose	SEER	
<i>Proprietary language models</i>												
GPT-4o-mini	75.81	67.58	47.80	83.01	83.79	66.85	54.00	59.14	65.24	73.91	54.54	66.52
GPT-3.5-turbo	59.31	58.12	37.40	74.11	71.11	56.22	39.05	57.84	67.85	86.96	73.61	61.96
<i>Open-source language models</i>												
Qwen2.5-7B	55.54	54.12	53.40	73.62	74.38	56.48	31.25	57.11	52.69	60.87	33.07	54.78
Llama3-8B	57.50	55.92	56.40	75.73	68.55	56.50	35.30	64.09	70.88	73.91	47.07	60.17
Llama3.1-8B	61.51	57.42	59.00	71.36	72.52	55.60	19.00	61.82	63.97	73.91	52.62	58.98
R1-Distill-Llama8B	50.12	48.89	46.60	70.55	68.42	53.49	36.10	55.73	62.04	69.57	31.71	53.93
QwQ-32B-preview	68.89	61.03	48.60	73.62	74.18	79.76	45.40	63.36	66.08	39.13	37.26	59.76
R1-Distill-Qwen32B	76.83	66.27	38.20	78.32	85.07	78.66	53.90	59.95	63.80	82.61	26.22	64.53
<i>Open-source medical models</i>												
MMedS-Ins	53.57	48.24	56.60	77.35	50.86	31.47	97.53	54.26	69.64	95.65	97.93	66.65
MedLlama3	55.85	59.36	66.40	84.63	70.08	47.97	22.50	62.39	68.10	69.57	50.69	59.78
Med42	50.20	49.70	55.40	74.76	61.43	57.26	31.35	59.14	81.57	65.22	37.14	56.65
OpenBioLLM	50.20	50.56	41.40	47.73	61.69	27.46	16.55	18.77	53.28	34.78	46.48	40.81
UltraMedical3-8B	68.89	61.82	51.60	80.58	75.08	45.18	36.70	66.13	72.73	60.87	24.55	58.56
UltraMedical3.1-8B	70.93	62.78	56.40	77.18	76.43	54.26	31.55	59.14	70.20	56.52	45.86	60.11
<i>Open-source slow-thinking medical models</i>												
HuatuoGPT-o1	62.53	59.31	69.20	87.70	70.53	50.98	40.20	24.61	66.08	56.52	46.85	57.68
MedS³ (Ours)												
CoT	65.91	60.55	56.80	78.48	75.66	55.84	51.65	57.03	64.73	73.91	48.97	62.68
SC	70.93	64.21	58.20	79.13	79.63	63.66	57.00	64.42	70.37	86.96	52.19	67.88
P-VS	71.88	65.20	59.60	80.10	79.50	77.12	65.20	73.03	79.97	95.65	58.36	73.24



Accelerating scientific breakthroughs with an AI co-scientist (Feb 19)

我们推出人工智能共同科学家 (AI co - scientist) ，这是一个以Gemini 2.0构建的多智能体人工智能系统，作为虚拟科学合作者来帮助科学家提出新颖的假设和研究提案，并加快科学与生物医学发现的进程。

受现代科学发现过程中未满足需求的激励，并基于近期人工智能（AI）的进展（包括能够跨复杂学科进行综合以及执行长期规划和推理的能力），我们开发了一个人工智能（AI）协科学家系统。

这个AI协科学家是一个多智能体AI系统，旨在成为科学家的协作工具。该AI协科学家系统基于Gemini 2.0构建，旨在反映支撑科学方法的推理过程。除了标准的文献综述、摘要和“深度研究”工具之外，AI协科学家系统旨在发现新的、原创性的知识，并根据先前的证据制定出可证明为新颖的研究假设和提案，且这些假设和提案要针对特定的研究目标。



Accelerating scientific breakthroughs with an AI co-scientist (Feb 19)

1、借助人工智能共同科学家赋能科学家并加速发现进程。

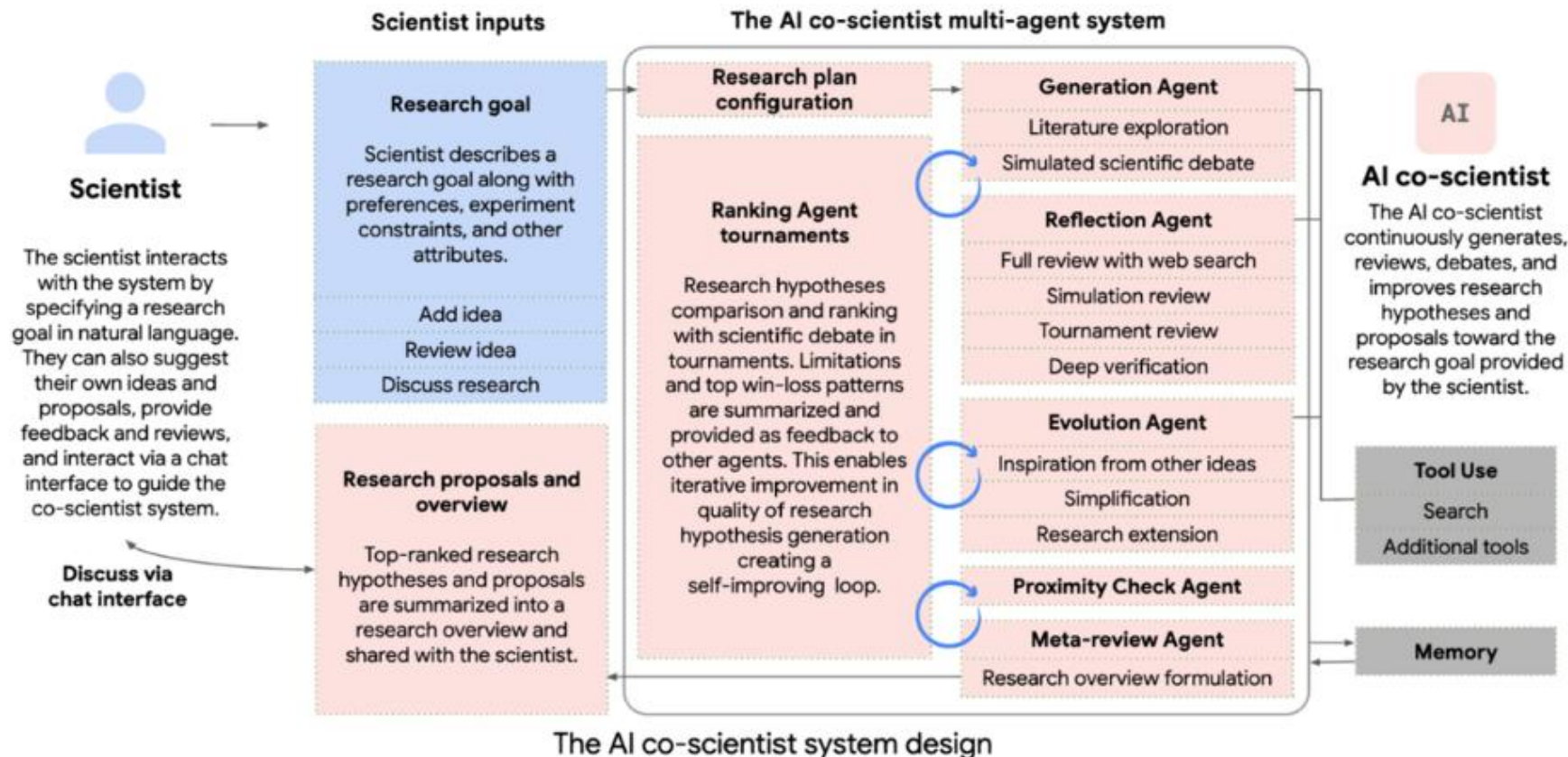
给定一个以自然语言明确表述的科学家研究目标，人工智能共同科学家旨在生成新颖的研究假设、详细的研究概述以及实验方案。为此，它运用了一个由专门代理组成的联盟——**生成 (Generation)**、**反思 (Reflection)**、**排序 (Ranking)**、**进化 (Evolution)**、**邻近性 (Proximity)** 和**元审查 (Meta-review)**——这些代理的灵感源于科学方法本身。这些代理利用自动化反馈来反复生成、评估和完善假设，从而形成一个自我改进的循环，不断产出质量越来越高且新颖的成果。

专为协作而设计，科学家可以通过多种方式与系统交互，包括直接提供自己的探索性种子想法，或以自然语言对生成的输出提供反馈。人工智能协科学家还利用网络搜索和专门的AI模型等工具，来增强生成假设的依据和质量。



Accelerating scientific breakthroughs with an AI co-scientist (Feb 19)

1、借助人工智能共同科学家赋能科学家并加速发现进程。



人工智能共同科学家将指定目标解析为一个研究计划配置，该配置由主管代理进行管理。主管代理将专业代理分配到工作队列，并分配资源。这种设计使系统能够灵活地扩展计算能力，并朝着特定的研究目标迭代改进其科学推理能力。



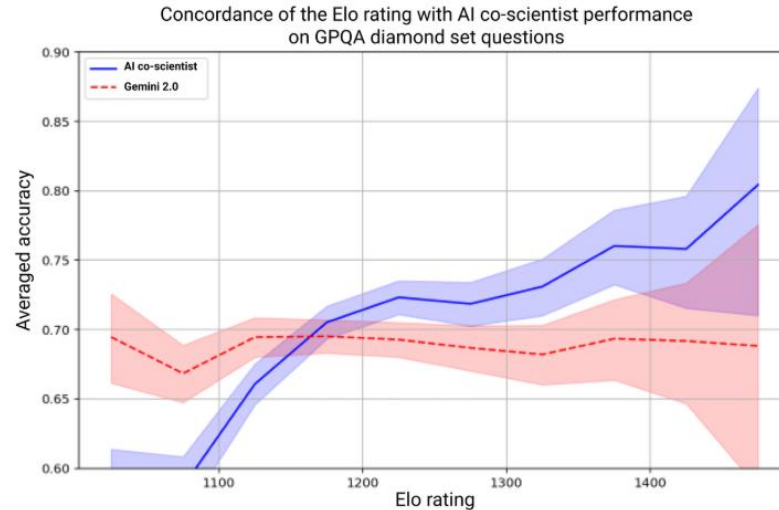
Accelerating scientific breakthroughs with an AI co-scientist (Feb 19)

2、在高级科学推理中扩展测试时计算量。

人工智能联合科学家利用**测试时计算资源扩展**来迭代地进行推理、演化和改进输出结果。

关键的推理步骤包括基于自我对弈的科学辩论以生成新假设、用于假设比较的排名竞赛，以及用于质量提升的“演化”过程。

该系统的智能体特性有助于递归式的自我批判，包括利用工具进行反馈以完善假设和提案。该系统的自我改进依赖于从其竞赛中得出的Elo自动评估指标。鉴于其核心作用，我们评估了更高的Elo评级是否与更高的输出质量相关。我们分析了Elo自动评级与该系统钻石集（一组具有挑战性的问题）上GPQA基准准确率之间的一致性，发现更高的Elo评级与更高正确答案概率呈正相关。





Accelerating scientific breakthroughs with an AI co-scientist (Feb 19)

3、用现实世界中的实验室实验验证新型人工智能共同科学家假设

Application	<i>Drug repurposing</i>	<i>Novel treatment target discovery</i>	<i>Explain mechanism of gene transfer evolution</i>
Challenge	Combinatorial search	Identifying novel targets	Understanding complex systems
Complexity	Medium	High	Very high
Scale	Moderate, data-limited	Moderate, experiment-limited	Large, data and computation-limited
Unknown elements	Constrained	Large	Vast and dynamic

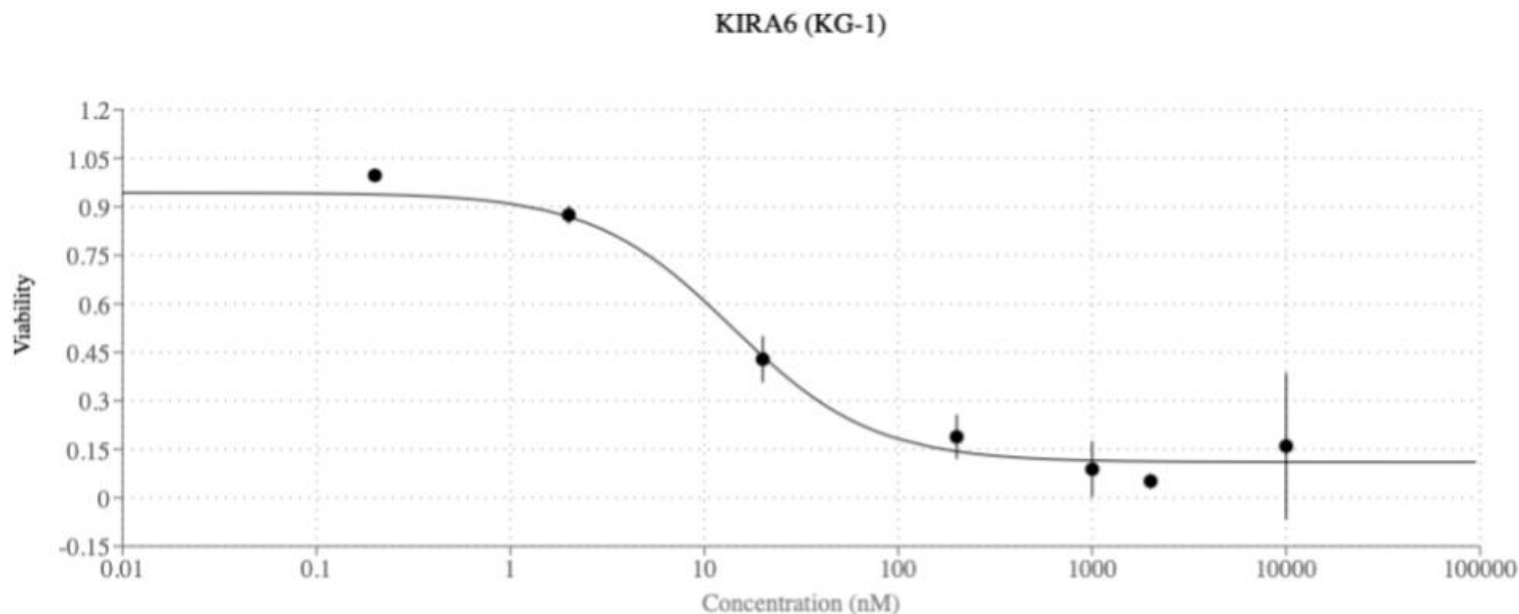


Accelerating scientific breakthroughs with an AI co-scientist (Feb 19)

3.1 急性髓系白血病的药物再利用。

药物开发是一个日益耗时且昂贵的过程，在这个过程中，对于每一种适应症或者疾病，新的治疗方法都需要重新开展发现和开发过程中的诸多环节。药物再利用则通过发掘现有药物超出其最初预期用途的新治疗应用来解决这一挑战。但是，由于这项任务较为复杂，它需要广泛的跨学科专业知识。

我们运用人工智能协科学家来协助预测药物再利用的机会，并且与我们的合作伙伴一起，通过计算生物学、专家临床医生的反馈以及体外实验对预测结果进行验证。值得注意的是，人工智能协科学家针对急性髓系白血病（AML）提出了新颖的再利用候选药物。随后的实验对这些提议进行了验证，证实所建议的药物在多种急性髓系白血病细胞系中能够在临床上相关的浓度下抑制肿瘤活力。



三种新型人工智能共同科学家预测的急性髓系白血病（AML）重新利用药物中的一种剂量 - 反应曲线。KIRA6在临床相关浓度下抑制KG - 1（急性髓系白血病细胞系）的活力。能够在较低的药物浓度下降低癌细胞的活力在多个方面具有优势，例如，因为它减少了脱靶副作用的可能性。

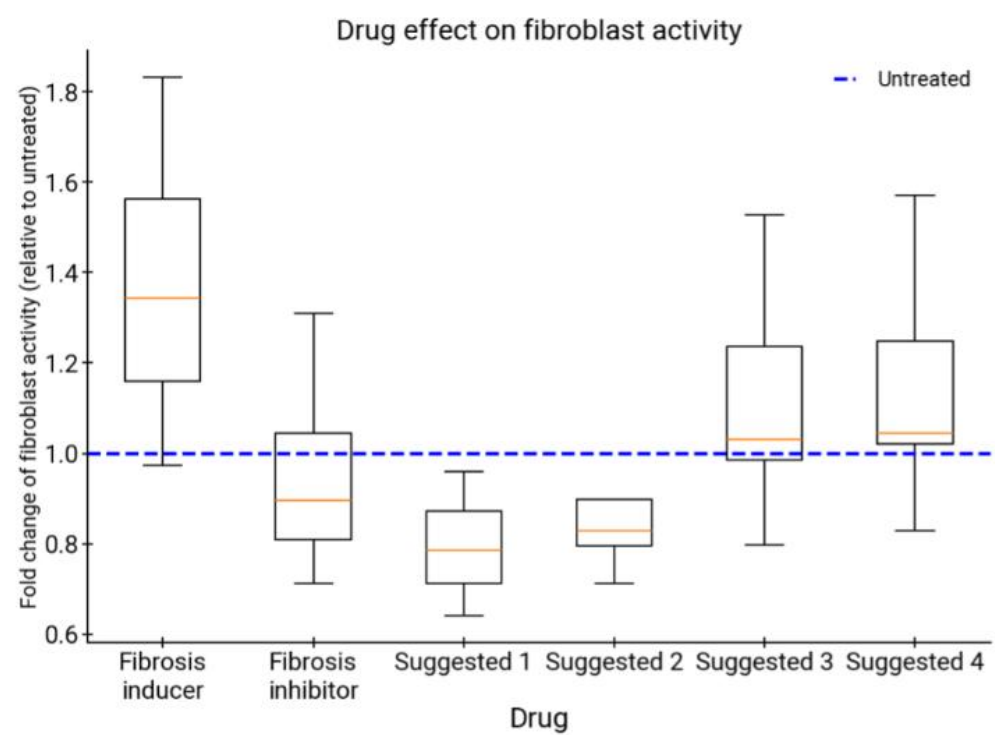


Accelerating scientific breakthroughs with an AI co-scientist (Feb 19)

3.2 推进肝纤维化的靶点发现

识别新的治疗靶点比药物再利用更为复杂，并且常常导致假设选择效率低下以及体外和体内实验优先级安排欠佳。人工智能辅助靶点发现有助于简化实验验证流程，可能有助于降低开发时间成本。

我们探究了人工智能共同科学家系统提出、排序并生成针对靶点发现假设的假设和实验方案的能力，重点关注肝纤维化。该人工智能共同科学家系统通过识别基于临床前证据且在对人类肝脏类器官（3D、多细胞组织培养物，由人类细胞衍生，旨在模拟人类肝脏的结构和功能）中有显著抗纤维化活性的表观遗传靶点，展现了其潜力。这些研究结果将在斯坦福大学合作者主导的一份即将发布的报告中详述。



与纤维化诱导剂（阴性对照）和抑制剂（阳性对照）相比，基于人工智能共同科学家建议的肝纤维化靶点的治疗比较。人工智能共同科学家建议的所有治疗方法均显示出良好的活性（所有建议药物的p值均 <0.01 ），包括可能逆转疾病表型的候选药物。相关结果详见斯坦福大学合作者即将发布的一份报告。

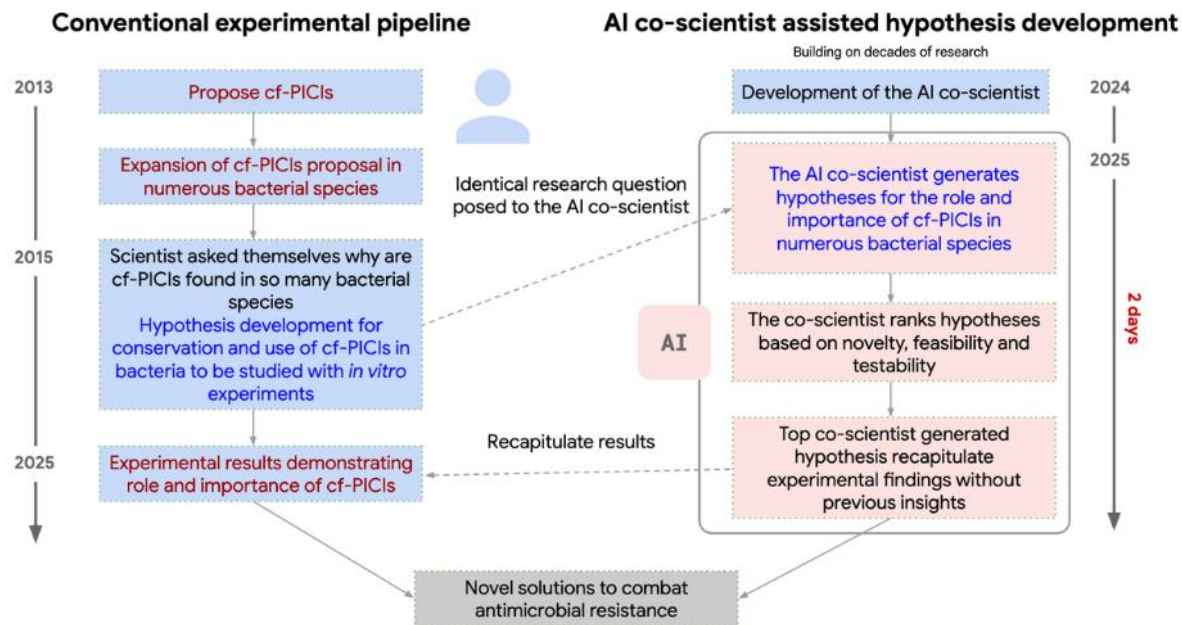


Accelerating scientific breakthroughs with an AI co-scientist (Feb 19)

3.3 解释抗菌素耐药性的机制

作为第三次验证，我们专注于生成假设来解释与抗菌素耐药性（AMR）相关的细菌基因转移进化机制——微生物进化出的抵抗治疗感染的药物的机制。这是另一个复杂的挑战，它涉及理解基因转移（接合、转导和转化）的分子机制以及驱动AMR基因传播的生态和进化压力。

在这次测试中，专家研究人员指示人工智能共同科学家探索一个在他们小组已经有过新发现，但尚未在公共领域揭示的主题，即解释形成衣壳的噬菌体诱导型染色体岛（cf - PICIs）如何存在于多种细菌物种中。人工智能共同科学家系统独立提出cf - PICIs通过与多种噬菌体尾部相互作用来扩大其宿主范围。这一计算机模拟发现，在使用人工智能共同科学家系统之前进行的原始新颖实验室实验中已经得到实验验证，并且与我们伦敦帝国理工学院和弗莱明倡议（Fleming Initiative）的合作者同时撰写的手稿（1， 2）中有所描述。这表明了人工智能共同科学家系统作为一种辅助技术的价值，因为它能够利用包含该主题所有先前开放获取文献在内的数十年研究成果。



人工智能（AI）共同科学家重新发现一种新型基因转移机制的时间线。蓝色：cf - PICIs动员发现的实验研究流程时间线。红色：人工智能（AI）共同科学家（在事先无相关知识的情况下）对这些关键发现的研发与重现。



Accelerating scientific breakthroughs with an AI co-scientist (Feb 19)

局限性与展望：

在我们的报告中，我们阐述了该系统的若干局限性以及改进的机会，其中包括加强文献综述、事实核查、与外部工具交叉核对、自动评估技术，以及开展涉及更多不同研究目标的学科专家的大规模评估。

人工智能联合科学家代表着朝着人工智能辅助技术方向的有前景的进步，这种技术可助力科学家加速发现进程。它能够在不同的科学和生物医学领域生成新颖且可检验的假设（其中一些已通过实验得到验证），并且随着计算能力的提升具备递归自我改进的能力，这表明它有潜力加速科学家解决科学和医学重大挑战的努力。我们期待对人工智能联合科学家作为科学家辅助工具的潜力进行负责任的探索。该项目展示了协作式、以人为中心的人工智能系统如何能够增强人类的创造力并加速科学发现。



Part **4**

总结



总结

从LLM的进展来看，目前的主要方向：

- 数据，数据，还是数据！！高质量的数据完全有可能超越大规模SFT的性能。
- 强化学习及其变体成为之后的主流方向。
- 小参数的大模型不再是梦想，包括MoE、蒸馏、量化等技术将成为重点。
- 复杂问题的推理能力，对Aha Moment更加深入的思考。
- 开源，开源，还是开源！

从LVLM的进展来看，目前的主要方向：

- 如何在VL中复现复杂推理能力和Aha Moment
- 如何在Fine-Gained的类别中实现更加出色的性能，超过专有的模型。
- 如何将强化学习的能力复制到VL中

从AI for Science的进展来看，目前的主要方向：

- Test-time Scaling
- 小数据，小算力，大模型
- 如同Co-Scientist一样更加复杂的系统
- 预训练的大模型
-



总结

LLM
LVLM
AI for Science

Where are we ?

技术层面上：

在各自研究的领域，处于技术金字塔的哪个阶段？

数据层面上：

在各自研究的领域，数据处于金字塔的哪个阶段？

算力层面上：

在各自的研究领域，算力处于什么阶段？



THANKS!