

01 研讨-第一季

2024.09.18

饮水思源 愛國榮校

刘鹏飞

GAIR: <https://plms.ai/>



内容规划

- O1的主要功能介绍
- 相关论文介绍
- 相关观点讨论
- Open问题讨论



内容规划

- O1的主要功能介绍
- 相关论文介绍**
- 相关观点讨论
- Open问题讨论



相关论文 (Part 1)

筛选	锁	类型	机构	日期
1	STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning	self-play	Stanford Google	2022/03/20
2	Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective	cot-theory	PKU	2023/05/24
3	Chain of Thought Empowers Transformers to Solve Inherently Serial Problems	cot-theory	Stanford Google	2024/02/20
4	Let's Verify Step by Step	process-reward	OpenAI	2023/05/31
5	Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking	internal-thought	Stanford Google	2024/03/14
6	Learn Beyond The Answer: Training Language Models with Reflection for Mathematical R...	internal-thought	Tencent	2024/06/17
7	Recursive Introspection: Teaching Language Model Agents How to Self-Improve	internal-thought	CMU UCB	2024/07/25
8	Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Para...	search	UCB Google	2024/08/06
9	Implicit Chain of Thought Reasoning via Knowledge Distillation	search-to-thought	AllenAI	2023/11/02
10	Grandmaster-Level Chess Without Search	search-to-thought	Google	2024/02/07



相关论文 (Part 1)

□ 🔒 A 论文

☰ 类型

☰ 机构

📅 日期

1	STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning	self-play	Stanford Google	2022/03/20
2	Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective	cot-theory	PKU	2023/05/24
3	Chain of Thought Empowers Transformers to Solve Inherently Serial Problems	cot-theory	Stanford Google	2024/02/20
4	Let's Verify Step by Step	process-reward	OpenAI	2023/05/31
5	Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking	internal-thought	Stanford Google	2024/03/14
6	Learn Beyond The Answer: Training Language Models with Reflection for Mathematical R...	internal-thought	Tencent	2024/06/17
7	Recursive Introspection: Teaching Language Model Agents How to Self-Improve	internal-thought	CMU UCB	2024/07/25
8	Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Para...	search	UCB Google	2024/08/06
9	Implicit Chain of Thought Reasoning via Knowledge Distillation	search-to-thought	AllenAI	2023/11/02
10	Grandmaster-Level Chess Without Search	search-to-thought	Google	2024/02/07



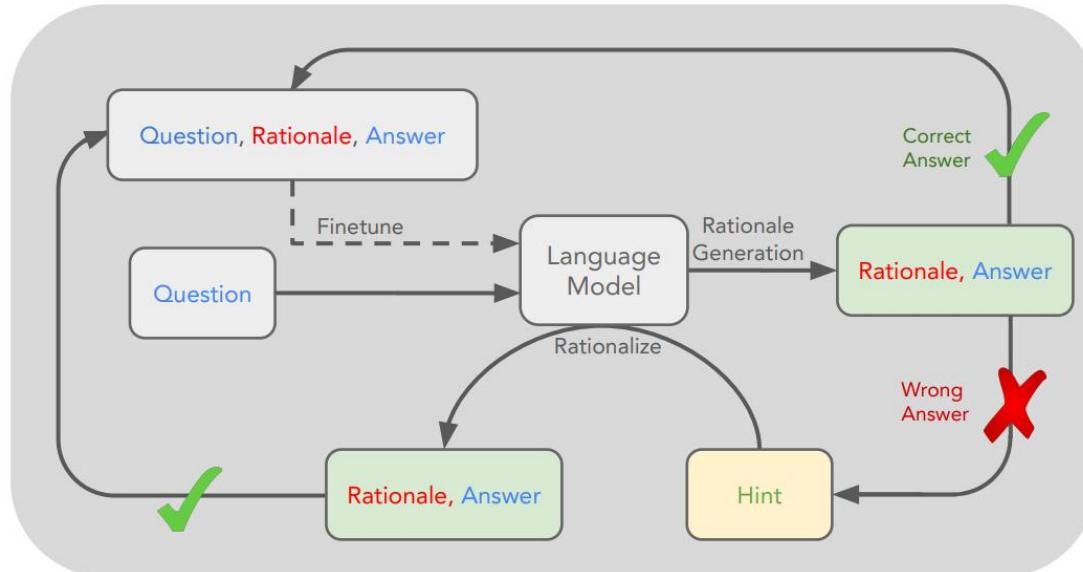
STaR: Self-Taught Reasoner (2022.03)

□ 核心贡献:

- 提出了一种无需检查新推理正确性的引导机制，生成推理数据集。
- 引入了“逆向推理”，通过提供正确答案帮助模型生成推理链。

□ 任务场景:

- 算术推理 (n位数加法)
- 常识问答
- 小学数学问题 (GSM8K)



Q: What can be used to carry a small dog?
Answer Choices:
(a) swimming pool
(b) basket
(c) dog show
(d) backyard
(e) own home
A: The answer must be something that can be used to carry a small dog. Baskets are designed to hold things. Therefore, the answer is basket (b).



STaR: Self-Taught Reasoner (2022.03)

□ 核心贡献:

- 提出了一种无需检查新推理正确性的引导机制，生成推理数据集。
- 引入了“逆向推理”，通过提供正确答案帮助模型生成推理链。

□ 任务场景:

- 算术推理 (n位数加法) :
- 常识问答
- 小学数学问题 (GSM8K)

1. 算术推理 (n位数加法)

任务：计算两个3位数的和。

- 输入: $624 + 259$
- 目标 (答案) : 883
- 生成的推理链:
 1. $624 + 259$, 记为C: 0 (进位为0)
 2. $4 + 9 = 13$, 记为3, 进位C: 1
 3. $2 + 5 + 1$ (进位) = 8, 记为8, 进位C: 0
 4. $6 + 2 = 8$, 记为8, 进位C: 0
 5. 最终答案: 883

最终答案: 883

推理链解读: 模型通过逐位计算，并考虑进位，得出最终正确答案883。

2. 常识问答 (CommonsenseQA)

任务：选择可以用来携带一只小狗的物品。

- 问题: 什么可以用来携带一只小狗?
 - (a) 游泳池
 - (b) 篮子
 - (c) 狗展
 - (d) 后院
 - (e) 自己的家
- 生成的推理链:

“答案应该是能用来携带小狗的物品。篮子通常是用来装东西的，因此答案是篮子 (b)。”

最终答案: 篮子 (b)

推理链解读: 模型通过常识推理，解释了篮子是携带小狗的合理选择。

3. 小学数学问题 (GSM8K)

任务：解决一个简单的数学问题，涉及多步计算。

- 问题: Lisa买了一件价值\$75的夹克，商店正在打折，给了她20%的折扣。她最终支付了多少?
- 生成的推理链:
 1. 原价是75美元，折扣是20%，因此折扣金额是 $75 * 0.20 = 15$ 美元。
 2. 最终支付金额是 $75 - 15 = 60$ 美元。

最终答案: 60美元

推理链解读: 模型通过分解问题，计算出折扣金额并得出最终支付的金额。



STaR: Self-Taught Reasoner (2022.03)

□ 任务场景：

- 算术推理 (n位数加法)：经过16次迭代后，模型在加法问题上的总准确率达到了89.5%，相比基线模型(76.3%)提升明显。
- 常识问答：最终共进行了16次迭代，达到了72.5%的准确率，与30倍大的GPT-3模型表现相近
- 小学数学问题 (GSM8K)：STaR方法共进行36次迭代（不包括逆向推理的迭代）。通过增加逆向推理，额外进行了10次迭代，从而使模型在测试集上的准确率提升到了10.7%

1. 算术推理 (n位数加法)

任务：计算两个3位数的和。

- 输入：624 + 259
- 目标（答案）：883
- 生成的推理链：
 1. $624 + 259$, 记为C: 0 (进位为0)
 2. $4 + 9 = 13$, 记为3, 进位C: 1
 3. $2 + 5 + 1$ (进位) = 8, 记为8, 进位C: 0
 4. $6 + 2 = 8$, 记为8, 进位C: 0
 5. 最终答案：883

最终答案：883
推理链解读：模型通过逐位计算，并考虑进位，得出最终正确答案883。

3. 小学数学问题 (GSM8K)

任务：解决一个简单的数学问题，涉及多步计算。

- 问题：Lisa买了一件价值\$75的夹克，商店正在打折，给了她20%的折扣。她最终支付了多少？
- 生成的推理链：
 1. 原价是75美元，折扣是20%，因此折扣金额是 $75 * 0.20 = 15$ 美元。
 2. 最终支付金额是 $75 - 15 = 60$ 美元。

最终答案：60美元
推理链解读：模型通过分解问题，计算出折扣金额并得出最终支付的金额。

2. 常识问答 (CommonsenseQA)

任务：选择可以用来携带一只小狗的物品。

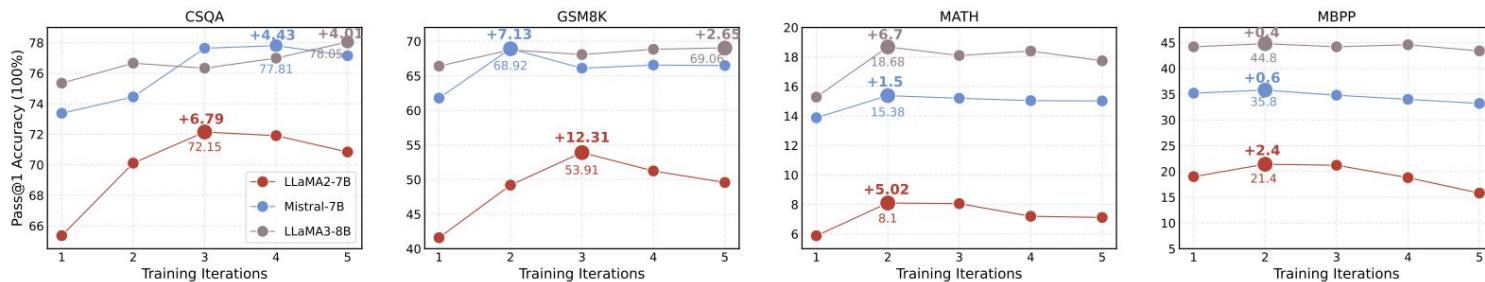
- 问题：什么可以用来携带一只小狗？
 - (a) 游泳池
 - (b) 篮子
 - (c) 狗展
 - (d) 后院
 - (e) 自己的家
- 生成的推理链：

“答案应该是能用来携带小狗的物品。篮子通常是用来装东西的，因此答案是篮子 (b)。”

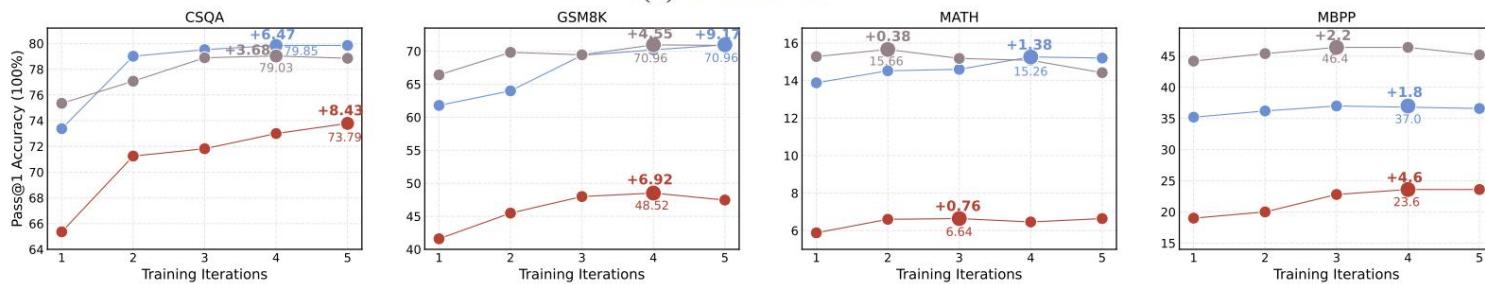
最终答案：篮子 (b)
推理链解读：模型通过常识推理，解释了篮子是携带小狗的合理选择。



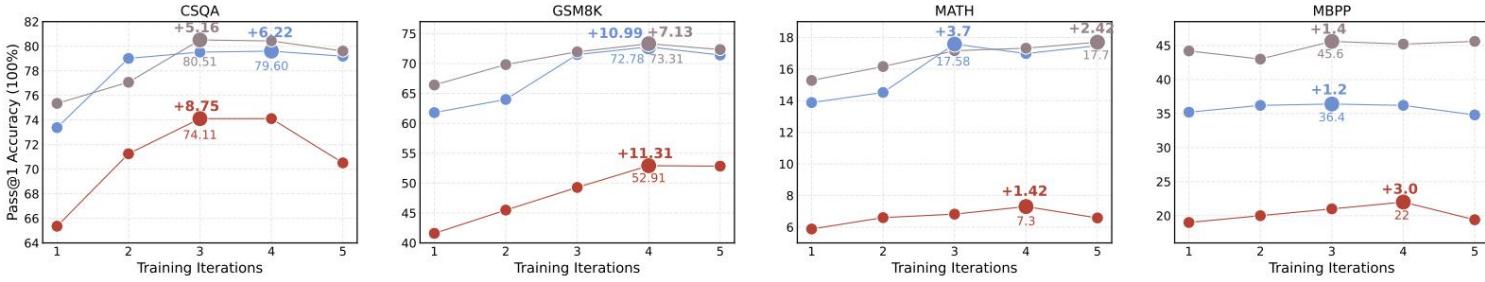
Self-Improvement Reversal in Post-training



(a) Iterative SFT



(b) Iterative DPO



(c) Iterative SFT-DPO

Progress or Regress? Self-Improvement Reversal in Post-training



相关论文 (Part 1)

筛选	锁定	类型	机构	日期
1	STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning	self-play	Stanford Google	2022/03/20
2	Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective	cot-theory	PKU	2023/05/24
3	Chain of Thought Empowers Transformers to Solve Inherently Serial Problems	cot-theory	Stanford Google	2024/02/20
4	Let's Verify Step by Step	process-reward	OpenAI	2023/05/31
5	Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking	internal-thought	Stanford Google	2024/03/14
6	Learn Beyond The Answer: Training Language Models with Reflection for Mathematical R...	internal-thought	Tencent	2024/06/17
7	Recursive Introspection: Teaching Language Model Agents How to Self-Improve	internal-thought	CMU UCB	2024/07/25
8	Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Para...	search	UCB Google	2024/08/06
9	Implicit Chain of Thought Reasoning via Knowledge Distillation	search-to-thought	AllenAI	2023/11/02
10	Grandmaster-Level Chess Without Search	search-to-thought	Google	2024/02/07



Mystery behind Chain of Thought (2023.05)

□ 核心贡献:

- 证明了有限深度的Transformers在没有CoT的情况下无法直接解决简单的算术或线性方程问题，除非其模型大小随着输入长度超多项式地增长。
- 当使用CoT提示时，即使是恒定大小的自回归Transformer也能够有效地解决这些问题。CoT通过分解问题为中间推理步骤，递归地增加了模型的“有效深度”，从而使其能够处理更复杂的问题。

任务类型	任务描述	CoT步骤数 (token数量)	复杂性因素
算术表达式求值	计算含有加法、减法、乘法和除法的数学表达式	4-6步 (取决于操作符数量)	操作符的数量
线性方程求解	求解多个变量的线性方程组	5-6步 (取决于变量数量)	方程和变量的数量
最长递增子序列 (LIS)	找出整数序列中最长的递增子序列	与输入序列长度相同 (如50至100个token)	输入序列的长度
编辑距离 (ED)	计算将一个字符串转换为另一个字符串的最少编辑距离	与字符串长度相同 (如12至20个token)	输入字符串的长度



Solve Inherently Serial Problems (2024.02)

核心贡献:

- 文章提出并证明了CoT可以显著提高Transformers在处理需要串行计算任务中的表现，而这类任务是平行计算难以处理的。
- 作者通过电路复杂性理论分析了没有CoT的Transformers受限于有限的并行计算能力（只能解决TC0和AC0类问题），而引入CoT后，Transformers可以通过多个推理步骤逐步计算复杂任务，解决常规Transformers难以处理的任务。
- 在嵌入大小为 $O(\log n)$ 的条件下，证明了CoT步数为多项式级别时，Transformers能够处理所有P/poly类问题。这表明嵌入大小和CoT步数共同决定了模型的表达能力。

任务类型	任务描述	CoT长度 (步数)	任务复杂性
模加法 (Modular Addition)	计算给定输入序列的模加和。例如，计算输入序列的和对一个特定数取模。	无需CoT，短序列也可处理 ($\log(n)$ 步)	可通过并行计算轻松解决，是可并行处理的问题 (TC0)
排列群组成 (Permutation Composition)	计算多个排列的复合，涉及多个元素之间的组合，类似于S5排列群的操作。	需要CoT (多步)	属于串行问题，复杂性较高 (不能通过简单并行计算解决)
迭代平方 (Iterated Squaring)	对一个数进行连续的平方运算，结果对一个模数取模。计算迭代次数越多，问题越复杂。	需要长CoT (随着迭代次数增加，CoT步数增加)	需要严格的串行计算，随着输入长度增长问题复杂度增加
电路值问题 (Circuit Value Problem, CVP)	给定一个布尔电路，计算电路的最终输出值。电路由多种逻辑门 (AND, OR, NOT等) 组成，要求按拓扑顺序评估每个门。	需要长CoT (随着电路深度增加，CoT步数增加)	严格的串行问题，属于P类问题，难以通过并行计算解决



相关论文 (Part 1)

筛选	锁	类型	机构	日期
1	STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning	self-play	Stanford Google	2022/03/20
2	Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective	cot-theory	PKU	2023/05/24
3	Chain of Thought Empowers Transformers to Solve Inherently Serial Problems	cot-theory	Stanford Google	2024/02/20
4	Let's Verify Step by Step	process-reward	OpenAI	2023/05/31
5	Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking	internal-thought	Stanford Google	2024/03/14
6	Learn Beyond The Answer: Training Language Models with Reflection for Mathematical R...	internal-thought	Tencent	2024/06/17
7	Recursive Introspection: Teaching Language Model Agents How to Self-Improve	internal-thought	CMU UCB	2024/07/25
8	Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Para...	search	UCB Google	2024/08/06
9	Implicit Chain of Thought Reasoning via Knowledge Distillation	search-to-thought	AllenAI	2023/11/02
10	Grandmaster-Level Chess Without Search	search-to-thought	Google	2024/02/07



Let's Verify Step by Step (2023.05)

The denominator of a fraction is 7 less than 3 times the numerator. If the fraction is equivalent to $\frac{2}{5}$, what is the numerator of the fraction? (Answer: 14)

Let's call the numerator x .

So the denominator is $3x-7$.

We know that $x/(3x-7) = 2/5$.

So $5x = 2(3x-7)$.

$5x = 6x - 14$.

So $x = 7$.

Figure 1: A screenshot of the interface used to collect feedback for each step in a solution.



Evaluating Mathematical Reasoning Beyond Accuracy

Question: Simplify $\sqrt{242}$

Ground Truth: $11\sqrt{2}$

Solution



ReasonEval

$S_{validity}$ $S_{redundancy}$

0.85 0.05

0.85 0.05

0.75 0.25

0.95 0.05

0.80 0.05

0.15 0.05

0.75 0.05

Step 1: We can simplify the square root of 242 by factoring it into its prime factors.

Step 2: The prime factorization of 242 is $2 * 11 * 11$.

Step 3: Let's check it. The prime factorization of 242 is $2 * 11 * 11$.

Step 4: By the property of square roots, we know that $\sqrt{ab} = \sqrt{a} * \sqrt{b}$ for any non-negative numbers a and b . Applying this property to $\sqrt{2 * 11 * 11}$, we have $\sqrt{2} * \sqrt{11^2}$.

Step 5: By the property of square roots, we know that $\sqrt{a^2} = a$ for any non-negative number a . Applying this property to $\sqrt{11^2}$, we have 11.

Step 6: By combining the results from step 4 and step 5, we have $\sqrt{2} * 11 * 11 = 11 * \sqrt{2}$.

Step 7: Therefore, the simplified form of $\sqrt{242}$ is $11\sqrt{2}$.

redundant invalid

Answer: $11\sqrt{2}$



Figure 1: Given a solution to a math problem, REASONEVAL scores each step and identifies the potential error location, serving as an extension to verify the final answer only.



相关论文 (Part 1)

论文	类型	机构	日期
1 STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning	self-play	Stanford Google	2022/03/20
2 Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective	cot-theory	PKU	2023/05/24
3 Chain of Thought Empowers Transformers to Solve Inherently Serial Problems	cot-theory	Stanford Google	2024/02/20
4 Let's Verify Step by Step	process-reward	OpenAI	2023/05/31
5 Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking	internal-thought	Stanford Google	2024/03/14
6 Learn Beyond The Answer: Training Language Models with Reflection for Mathematical R...	internal-thought	Tencent	2024/06/17
7 Recursive Introspection: Teaching Language Model Agents How to Self-Improve	internal-thought	CMU UCB	2024/07/25
8 Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Para...	search	UCB Google	2024/08/06
9 Implicit Chain of Thought Reasoning via Knowledge Distillation	search-to-thought	AllenAI	2023/11/02
10 Grandmaster-Level Chess Without Search	search-to-thought	Google	2024/02/07



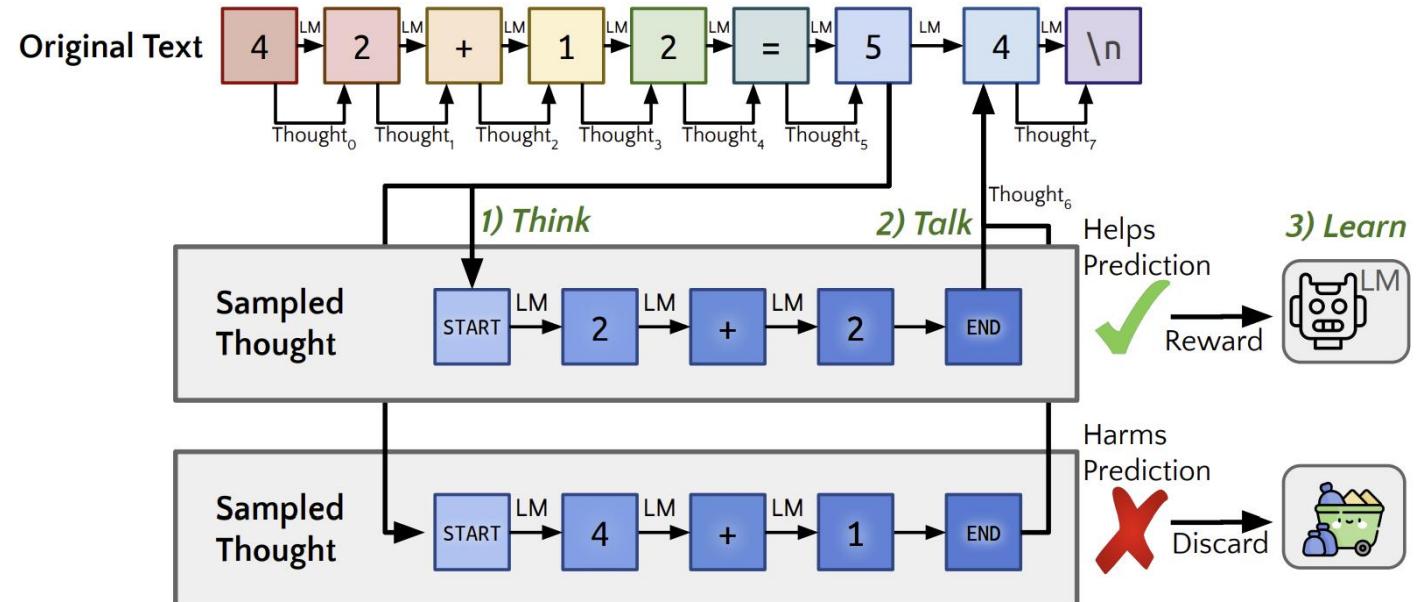
Quiet-STaR (2024.03)

核心贡献:

- 并不是在显式地“说出”这些推理过程，而是模型在内部“静默”地生成并使用这些推理链来帮助预测下一个词。
- 技术细节贡献:
 - 并行推理生成
 - 非短视损失优化
 - REINFORCE算法的使用

任务场景(Mistral CPT)

- 常识问答(5-10)
- GSM8K(10-20)



```
'<s> # Magnesium reacts with nitrogen to form magnesium nitride. The chemical formula for this reaction is Mg+N_2-> MgN_2. What is the product, or what are the products, of this reaction?\n\nJan 12, 2016\n\nThe formula for magnesium nitride is $M\{g\}_3\{N\}_2$.\n\n#### Explanation:\nAs do many active metals, magnesium nitride can be<|startofthought|> 1 --, so the equation of the reaction that forms magnesium nitride is\n\n$Mg + N_2 \\rightarrow<|endofthought|> formed by heating the metal (fier'
```



Learn Beyond The Answer (2024.06)

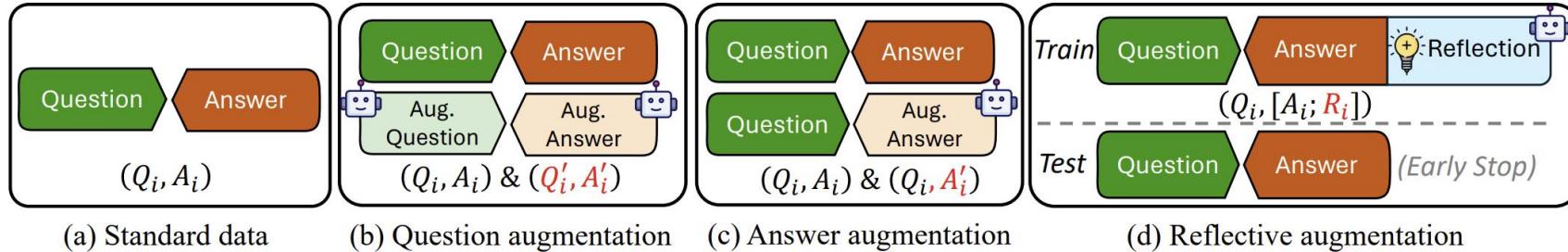
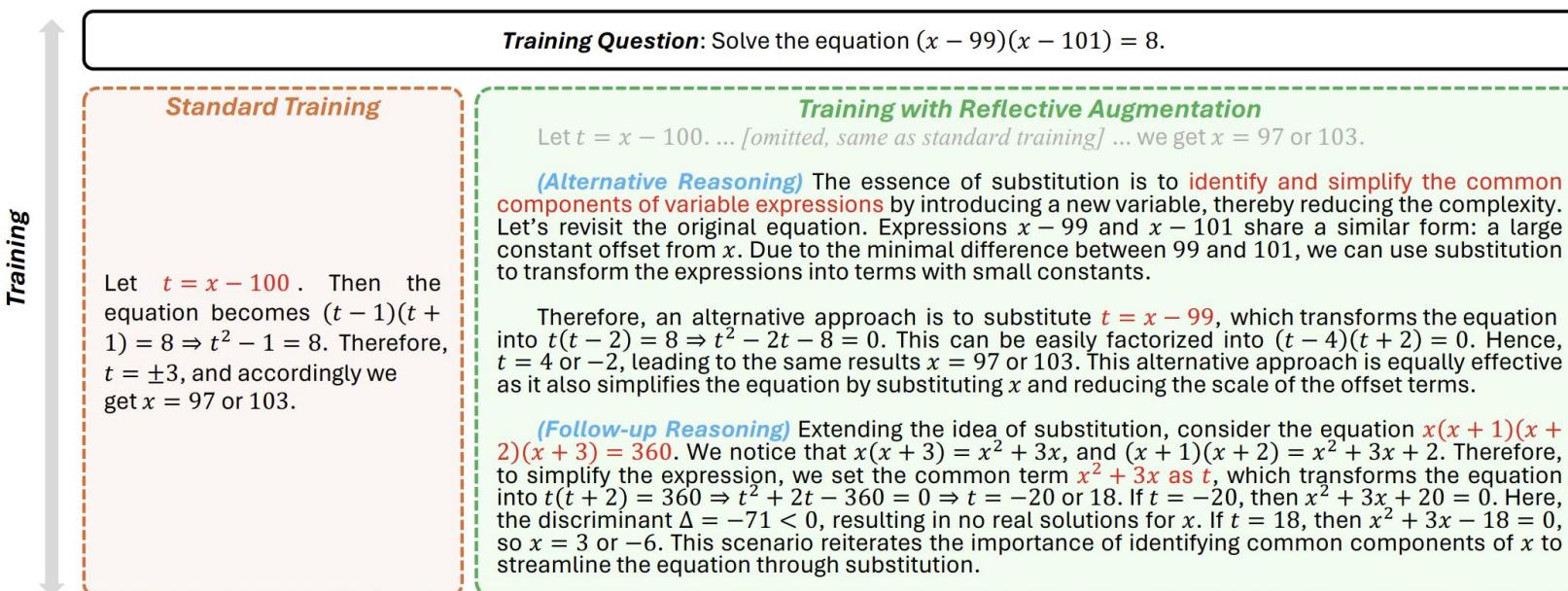


Figure 1: Question augmentation creates new questions based on existing ones. Answer augmentation re-samples answers for each problem to increase diversity. Both methods expand the size of the training set. Reflective augmentation appends the original answer with a **reflective section**, which is complementary to traditional approaches. Corresponding training sequences are shown in an (input, output) format, where augmented parts are in red.





Learn Beyond The Answer (2024.06)

Model	Training Data	In-Distribution		Out-Of-Distribution					Avg.
		GSM	MATH	Mathematics	MAWPS	SVAMP	MMLU-Math	SAT-Math	
<i>Standard Training Data</i>									
Mistral	Standard	56.25	13.96	14.80	73.07	53.50	37.68	31.82	40.15
	Standard + RefAug	60.05	17.36	19.40	80.25	59.30	43.63	48.64	46.95
Gemma	Standard	60.05	17.06	19.80	76.81	57.10	39.32	42.73	44.70
	Standard + RefAug	64.59	23.04	26.70	85.64	64.70	46.61	55.00	52.33
<i>Question Augmentation Data</i>									
Mistral	Q-Aug	56.03	18.06	18.00	79.99	59.10	38.19	36.16	43.65
	Q-Aug×2	59.14	21.26	20.90	80.84	61.50	40.86	46.82	47.33
	Q-Aug + RefAug	63.00	21.66	20.50	81.78	60.20	42.20	50.91	48.61
Gemma	Q-Aug	61.11	21.98	23.90	81.78	59.70	40.45	48.18	48.16
	Q-Aug×2	63.68	24.42	23.50	82.12	59.50	42.71	48.18	49.16
	Q-Aug + RefAug	68.61	26.38	28.70	85.39	66.00	48.05	51.82	53.56
<i>Answer Augmentation Data</i>									
Mistral	A-Aug	66.19	23.08	23.90	81.10	62.20	37.78	40.91	47.88
	A-Aug×2	67.93	27.12	28.30	83.26	66.50	42.61	45.91	51.66
	A-Aug + Q-Aug	69.67	24.32	26.90	81.82	61.20	38.50	46.82	49.90
	A-Aug + RefAug	72.93	29.40	31.20	84.41	71.50	47.74	60.45	56.80
Gemma	A-Aug	68.31	28.78	33.10	83.05	65.10	46.51	61.36	55.17
	A-Aug×2	70.66	31.14	33.30	85.22	69.70	47.13	54.55	55.96
	A-Aug + RefAug	74.15	33.60	38.20	85.68	69.10	52.26	64.09	59.58
<i>MetaMath Augmentation Data</i>									
Mistral	MetaMath _{40k}	68.46	20.96	20.30	85.09	66.50	38.09	42.73	48.88
	MetaMath _{80k}	69.29	23.54	23.20	86.75	68.60	41.17	43.64	50.88
	MetaMath _{40k} + RefAug _{40k}	73.84	26.60	27.00	87.68	75.30	44.15	53.18	55.39
	MetaMath _{400k} *	77.48	28.42	33.00	90.10	79.10	48.77	55.00	58.84
	MetaMath _{400k} + RefAug _{40k}	78.70	32.50	34.50	91.59	77.90	49.69	59.09	60.57
	MetaMath _{400k} (CT)	78.39	28.72	32.70	90.87	78.90	49.08	55.91	59.22
<i>MetaMath Augmentation Data (CT)</i>									
Mistral	MetaMath _{400k} + RefAug _{40k} (CT)	78.92	30.12	36.20	91.46	79.90	49.69	57.27	60.51



Recursive Introspection (2024.07)

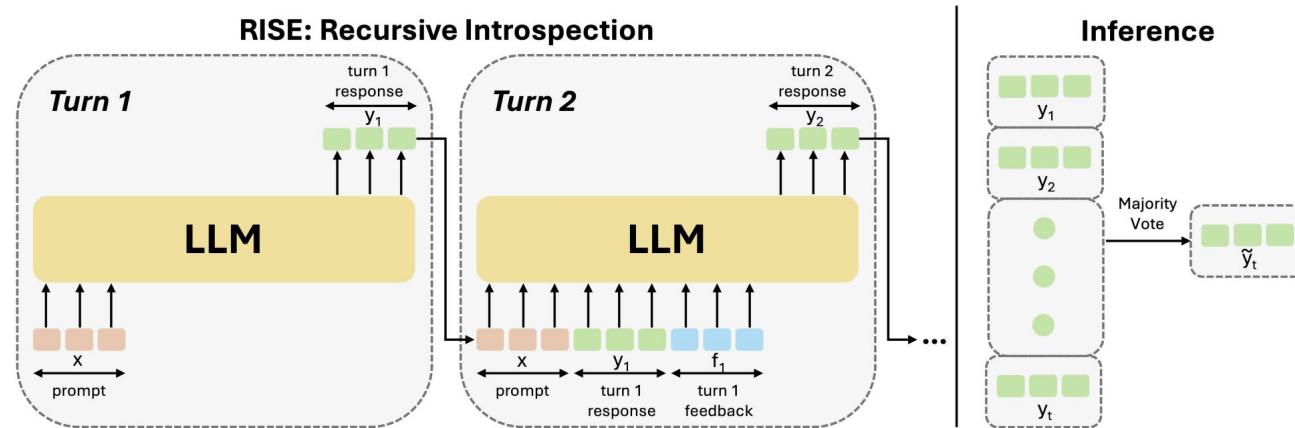
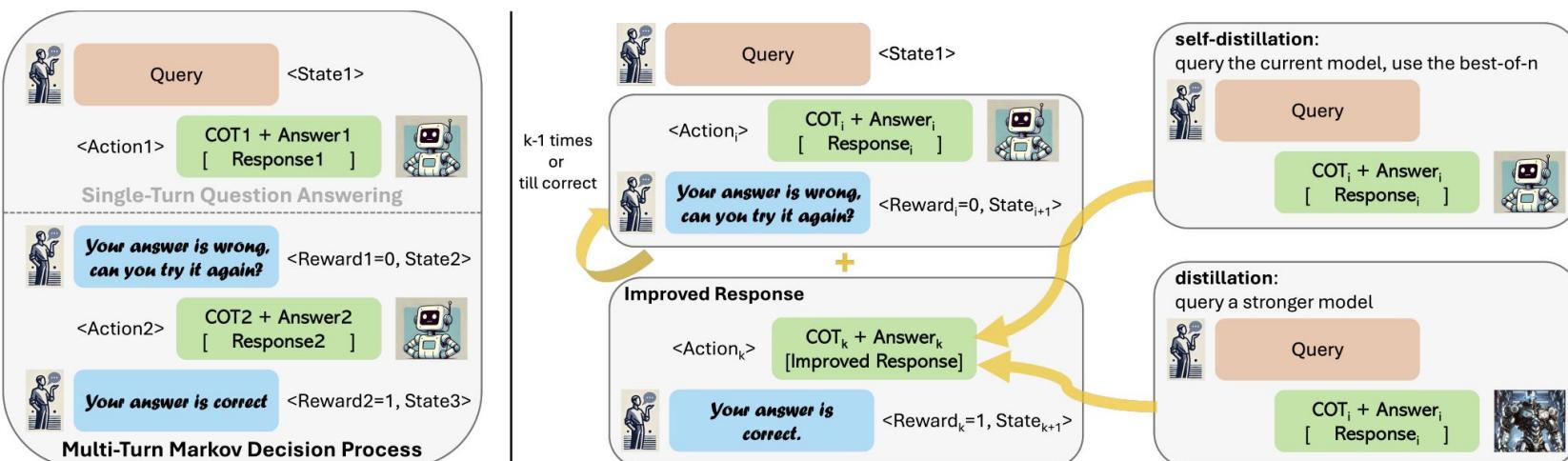


Figure 1: Recursive Introspection (RISE). Using iterative multi-round training on on-policy rollouts and supervision from a reward function, RISE trains models that are capable of improving themselves over multiple turns. At inference, we run majority voting on candidate outputs from different turns to obtain the final response.



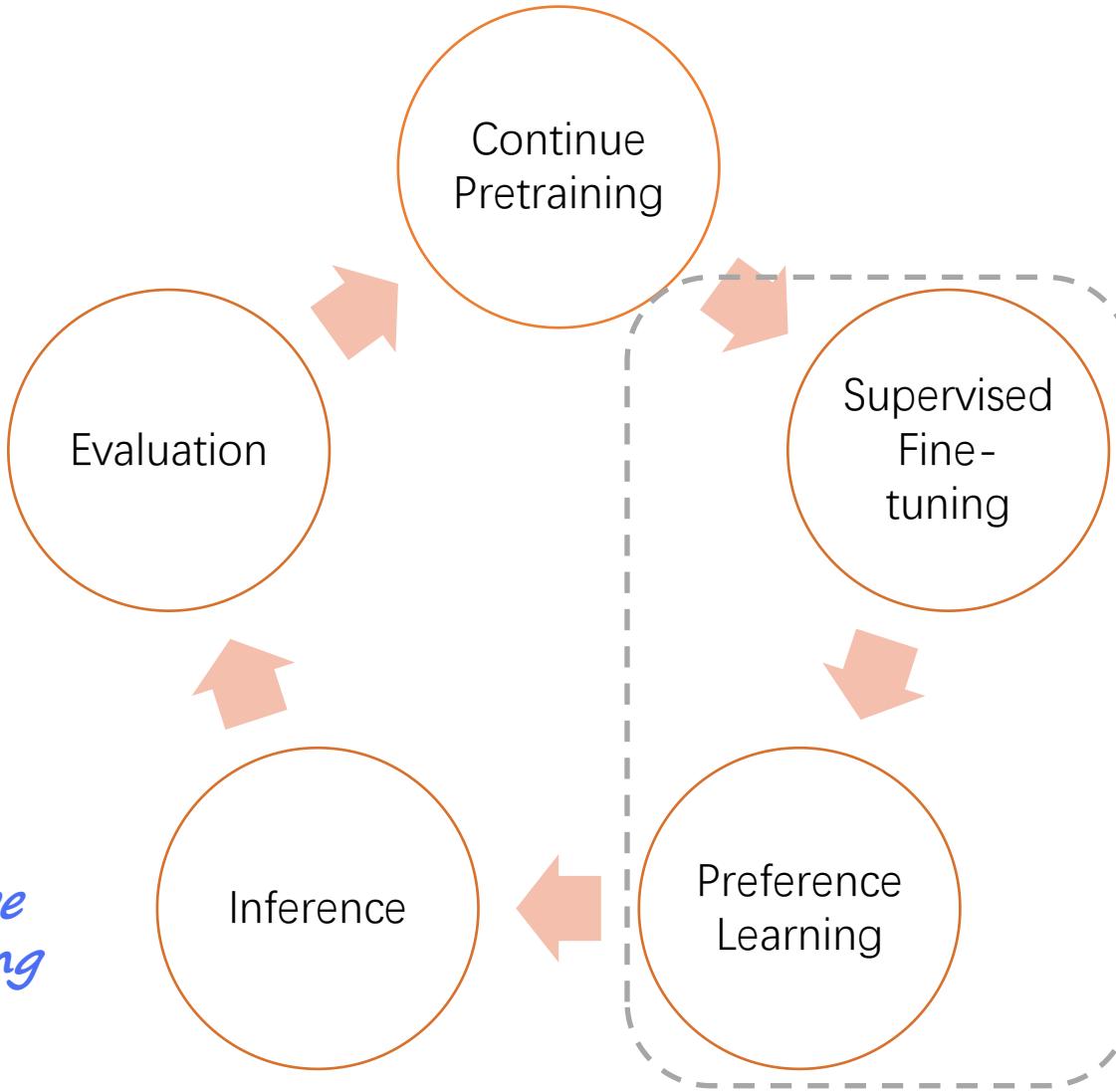


相关论文 (Part 1)

论文	类型	机构	日期
1 STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning	self-play	Stanford Google	2022/03/20
2 Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective	cot-theory	PKU	2023/05/24
3 Chain of Thought Empowers Transformers to Solve Inherently Serial Problems	cot-theory	Stanford Google	2024/02/20
4 Let's Verify Step by Step	process-reward	OpenAI	2023/05/31
5 Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking	internal-thought	Stanford Google	2024/03/14
6 Learn Beyond The Answer: Training Language Models with Reflection for Mathematical R...	internal-thought	Tencent	2024/06/17
7 Recursive Introspection: Teaching Language Model Agents How to Self-Improve	internal-thought	CMU UCB	2024/07/25
8 Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Para...	search	UCB Google	2024/08/06
9 Implicit Chain of Thought Reasoning via Knowledge Distillation	search-to-thought	AllenAI	2023/11/02
10 Grandmaster-Level Chess Without Search	search-to-thought	Google	2024/02/07



Which to Scale?



“Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters” - Google DeepMind

“我们在 post-training 环节有很多进步，未来我们也会继续延用这个技术，而且有可能增加在这方面投入的计算资源”

- John Schulman



Scaling LLM Test-Time Compute (2024.08)

□ 核心贡献:

- 测试了两种主要机制
 - 基于验证器模型的搜索 (Verifier Search)
 - 自适应更新模型的分布
- 实验使用了MATH基准测试，结果表明，较小的模型在推理阶段使用更多的计算资源时，其性能可以超越比它大14倍的模型



相关论文 (Part 1)

□	🔒 A 论文	☰ 类型	☰ 机构	📅 日期
1	STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning	self-play	Stanford Google	2022/03/20
2	Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective	cot-theory	PKU	2023/05/24
3	Chain of Thought Empowers Transformers to Solve Inherently Serial Problems	cot-theory	Stanford Google	2024/02/20
4	Let's Verify Step by Step	process-reward	OpenAI	2023/05/31
5	Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking	internal-thought	Stanford Google	2024/03/14
6	Learn Beyond The Answer: Training Language Models with Reflection for Mathematical R...	internal-thought	Tencent	2024/06/17
7	Recursive Introspection: Teaching Language Model Agents How to Self-Improve	internal-thought	CMU UCB	2024/07/25
8	Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Para...	search	UCB Google	2024/08/06
9	Implicit Chain of Thought Reasoning via Knowledge Distillation	search-to-thought	AllenAI	2023/11/02
10	Grandmaster-Level Chess Without Search	search-to-thought	Google	2024/02/07



Implicit CoT (2023. 11)

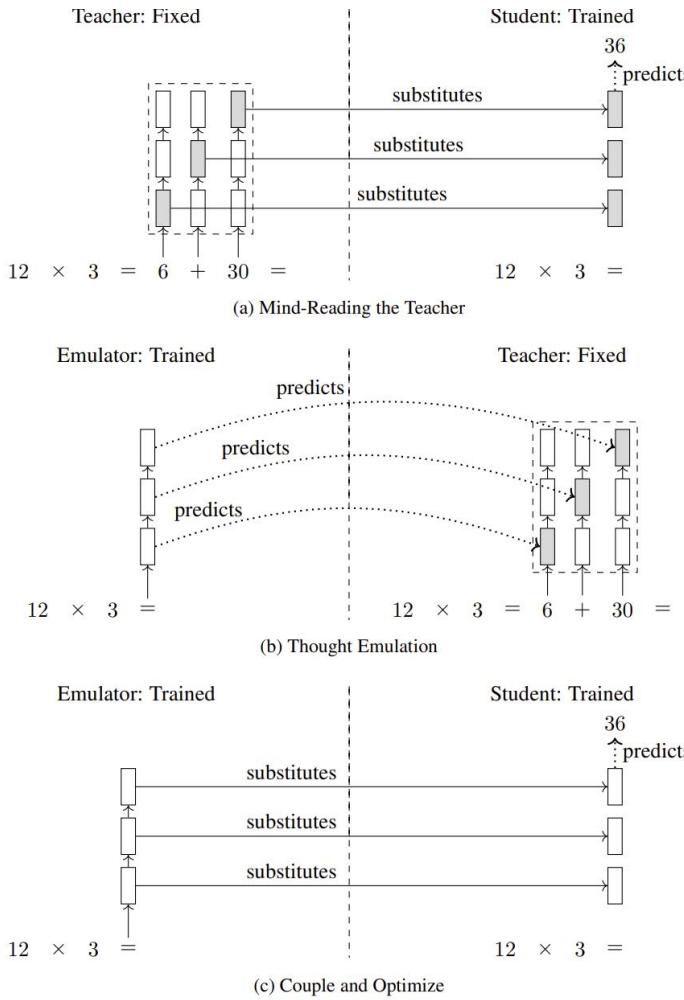


Table 3: Main results. Accuracy (Acc) measures the exact match accuracy of producing the final answer. Throughput measures the number of examples per second during inference using a batch size of 1, and is normalized by the throughput of the corresponding No CoT model.

Model	#Layers	4 × 4 Mult		5 × 5 Mult		GSM8K-Aug	
		Acc	Throughput	Acc	Throughput	Acc	Throughput
No CoT							
GPT-2 Small	12	0.29	1.00	0.01	1.00	0.13	1.00
GPT-2 Medium	24	0.76	1.00	0.02	1.00	0.17	1.00
GPT-2 Large	36	0.34	1.00	0.01	1.00	0.13	1.00
Implicit CoT							
GPT-2 Small	12	0.97	0.67	0.10	0.71	0.20	0.66
GPT-2 Medium	24	0.96	0.69	0.96	0.73	0.22	0.66
Explicit CoT							
GPT-2 Small	12	1.00	0.17	1.00	0.14	0.41	0.08
GPT-2 Medium	24	1.00	0.17	1.00	0.14	0.44	0.08
GPT-2 Large	36	1.00	0.17	0.99	0.15	0.45	0.08



Grandmaster-Level Chess Without Search (2024.02)

The recent breakthrough successes in machine learning are mainly attributed to scale: namely large-scale attention-based architectures and datasets of unprecedented scale. This paper investigates the impact of training at scale for chess. Unlike traditional chess engines that rely on complex heuristics, explicit search, or a combination of both, we train a 270M parameter transformer model with supervised learning on a dataset of 10 million chess games. We annotate each board in the dataset with action-values provided by the powerful Stockfish 16 engine, leading to roughly 15 billion data points. Our largest model reaches a Lichess blitz Elo of 2895 against humans, and successfully solves a series of challenging chess puzzles, without any domain-specific tweaks or explicit search algorithms. We also show that our model outperforms AlphaZero's policy and value networks (without MCTS) and GPT-3.5-turbo-instruct. A systematic investigation of model and dataset size shows that strong chess performance only arises at sufficient scale. To validate our results, we perform an extensive series of ablations of design choices and hyperparameters.



相关论文 (Part 1)

论文	类型	机构	日期
1 STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning	self-play	Stanford Google	2022/03/20
2 Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective	cot-theory	PKU	2023/05/24
3 Chain of Thought Empowers Transformers to Solve Inherently Serial Problems	cot-theory	Stanford Google	2024/02/20
4 Let's Verify Step by Step	process-reward	OpenAI	2023/05/31
5 Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking	internal-thought	Stanford Google	2024/03/14
6 Learn Beyond The Answer: Training Language Models with Reflection for Mathematical R...	internal-thought	Tencent	2024/06/17
7 Recursive Introspection: Teaching Language Model Agents How to Self-Improve	internal-thought	CMU UCB	2024/07/25
8 Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Para...	search	UCB Google	2024/08/06
9 Implicit Chain of Thought Reasoning via Knowledge Distillation	search-to-thought	AllenAI	2023/11/02
10 Grandmaster-Level Chess Without Search	search-to-thought	Google	2024/02/07



内容规划

- O1的主要功能介绍
- 相关论文介绍
- **相关观点讨论**
- Open问题讨论



内容规划

- O1的主要功能介绍
- 相关论文介绍
- 相关观点讨论
- Open问题讨论



分析下o1 thought的规律?

- 有没有什么stopping symbol (通过分析高频的词语)
- 平均长度 或者长度和 题目 (难度) 的关系
- 思考reward model怎么设计



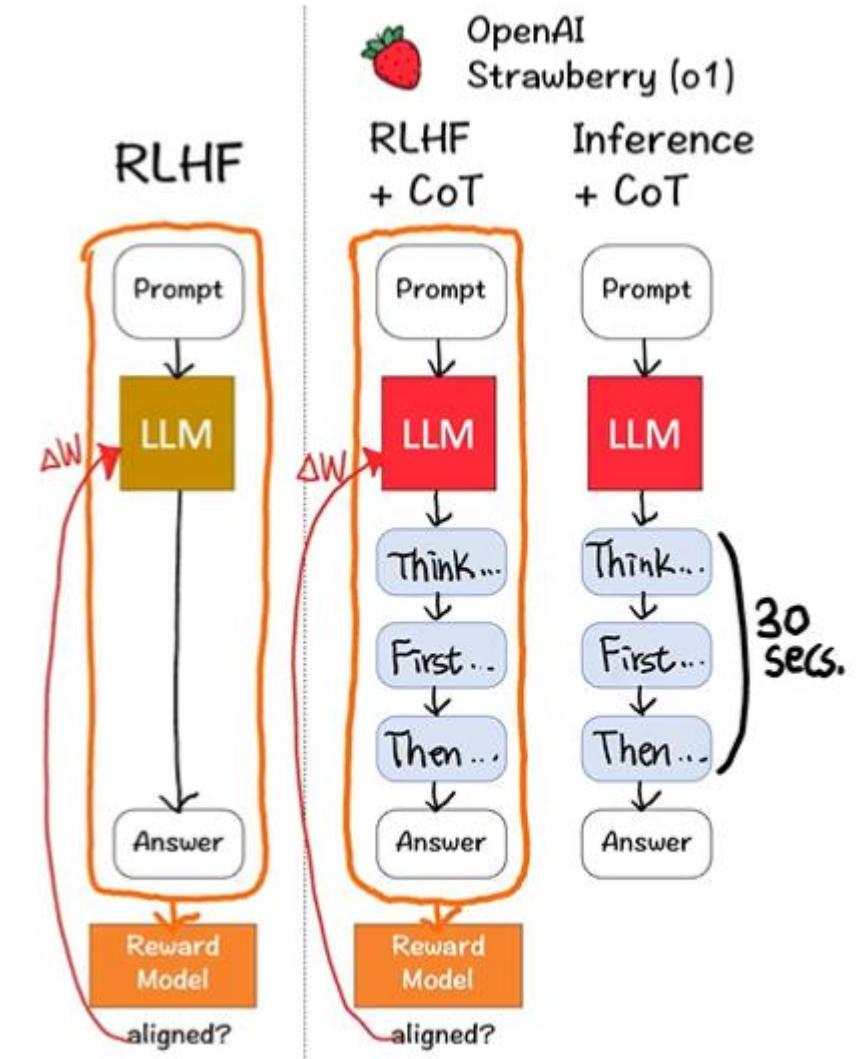
COT里面都能放哪些信息？

- OpenAI: Through training, the models learn to **refine their thinking process, try different strategies, and recognize their mistakes**
- Others?



如何Scaling COT?

- How many samples ? How to annotate/synthesize samples?
- Pretraining stage? Only SFT? - RLHF? Reward Model怎么用?





Q1还有哪些问题解决不了?

□ 负责逻辑?



Q1: Has o1 reached level 2?

Negative

Introduction

Question 1

Question 2

Limitations

Future Work

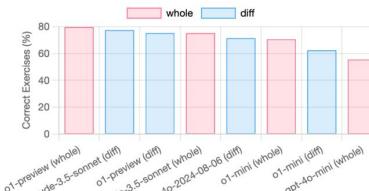
How many 'r's are there in the word strawberry?

思考一秒
在单词 "strawberry" 中, 有 两个 字母 "r".

具体分解如下:

- s
- t
- r
- a
- w
- b
- e
- r
- r
- y

因此, "strawberry" 中包含 两个 "r".



aider

hello

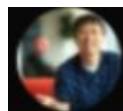
思考 5 秒

Hello! How can I assist you today?

The increasing test time make the comparison unfair

[Other evaluation results](#)

6 not reach the level of a decent PhD student who can think from first principles and coming up with novel ideas and patterns?



Yi Ma @YiMaTweets · 11h

9.11 is still larger than 9.8, despite can memorize solutions to PhD level questions. Again, memorizing is not understanding and knowledge is not intelligence.

26

46

344

41K



ive, for math problems that require out-of-box es poorly. Surprised to see o1-preview cannot even us errors in its logic chain. See below for a few

hat it shows in o1-preview) assumes "there is always an en problem and is trained to match existing known e a "small-town swot"?



01 背景下，我们有哪些新研究/工程的问题可以做？

- 为什么long cot可以泛化这么好？

谢谢各位！