

rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking

Harald Semmelrock



rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking

Xinyu Guan* Li Lyna Zhang**[◇] Yifei Liu
Ning Shang Youran Sun Yi Zhu Fan Yang Mao Yang

Microsoft Research Asia

Abstract

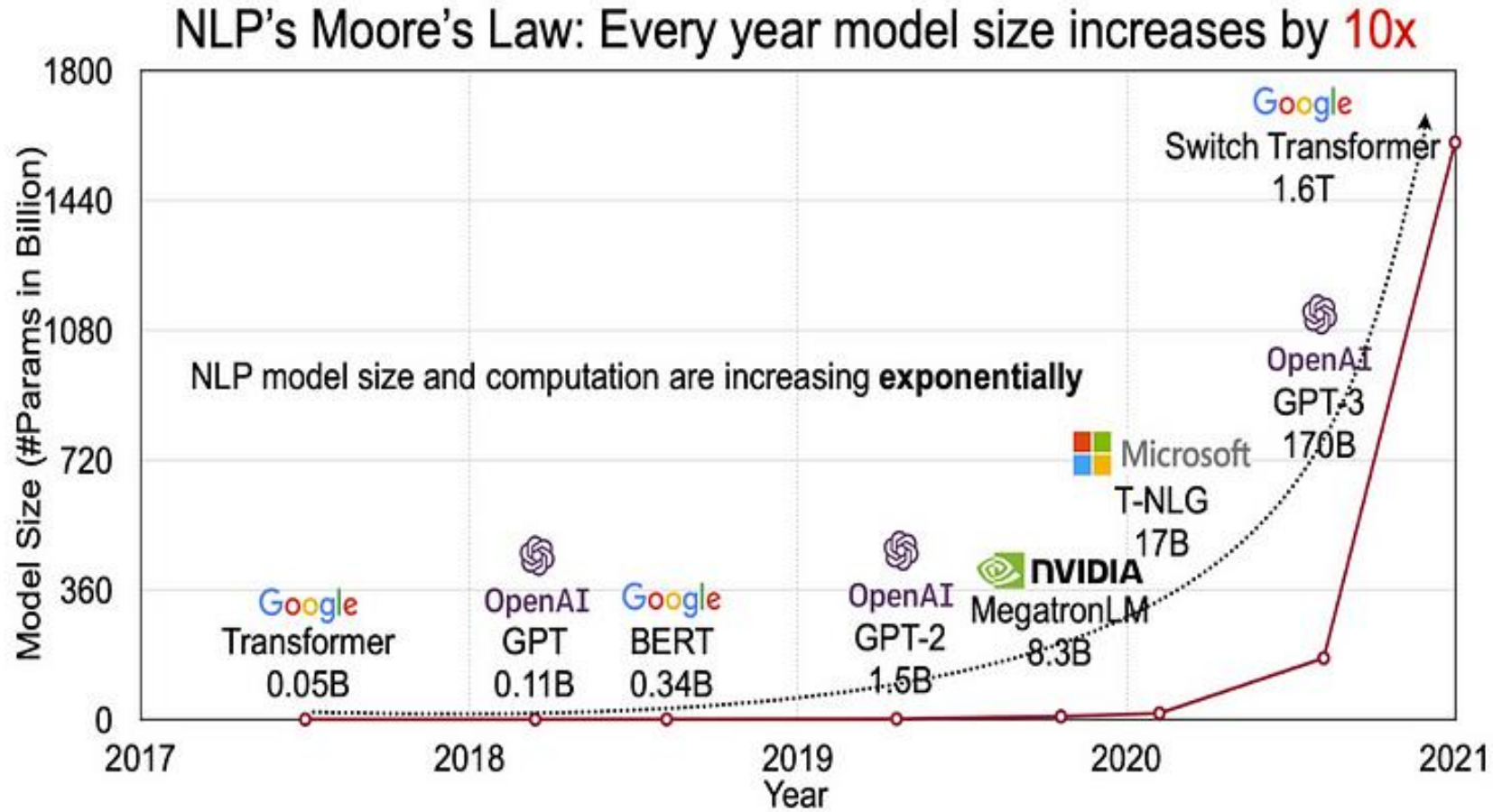
We present rStar-Math to demonstrate that small language models (SLMs) can rival or even surpass the math reasoning capability of OpenAI o1, without distillation from superior models. rStar-Math achieves this by exercising “deep thinking” through Monte Carlo Tree Search (MCTS), where a math *policy SLM* performs test-time search guided by an SLM-based *process reward model*. rStar-Math introduces three innovations to tackle the challenges in training the two SLMs: **(1)** a novel code-augmented CoT data synthesis method, which performs extensive MCTS rollouts to generate *step-by-step verified reasoning trajectories* used to train the policy SLM; **(2)** a novel process reward model training method that avoids naïve step-level score annotation, yielding a more effective *process preference model (PPM)*; **(3)** a *self-evolution recipe* in which the policy SLM and PPM are built from scratch and iteratively evolved to improve reasoning capabilities. Through 4 rounds of self-evolution with millions of synthesized solutions for 747k math problems, rStar-Math boosts SLMs’ math reasoning to state-of-the-art levels. On the MATH benchmark, it improves Qwen2.5-Math-7B from 58.8% to 90.0% and Phi3-mini-3.8B from 41.4% to 86.4%, surpassing o1-preview by +4.5% and +0.9%. On the USA Math Olympiad (AIME), rStar-Math solves an average of 53.3% (8/15) of problems, ranking among the top 20% the brightest high school math students. Code and data will be available at <https://github.com/microsoft/rStar>.

Task (pass@1 Acc)	rStar-Math (Qwen-7B)	rStar-Math (Qwen-1.5B)	rStar-Math (Phi3-mini)	OpenAI o1-preview	OpenAI o1-mini	QWQ 32B-preview	GPT-4o	DeepSeek-V3
MATH	90.0	88.6	86.4	85.5	90.0	90.6	76.6	90.2
AIME 2024	53.3	46.7	43.3	44.6	56.7	50.0	9.3	39.2
Olympiad Bench	65.6	64.6	60.3	-	65.3	61.2	43.3	55.4
College Math	60.5	59.3	59.1	-	57.8	55.8	48.5	58.9
Omni-Math	50.5	48.5	46.0	52.5	60.5	49.6	30.5	35.9

Table 1: rStar-Math enables frontier math reasoning in SLMs via deep thinking over 64 trajectories.

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Traditional LLM scaling



The Increasing size of LLMs, Medium, 15.07.2023, Retrieved from:
<https://medium.com/@harishdatalab/unveiling-the-power-of-large-language-models-llms-e235c4eba8a9>

Hallucinations



Who was the first person to walk on the moon?



Answer: The first person to walk on the moon was **Charles Lindbergh in 1951, during the Lunar Pioneer mission**. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. ❌



Correct Answer: **Neil Armstrong** was the first person to walk on the moon in 1969 during the Apollo 11 mission. ✅

(a) Factuality Hallucination



Please summarize the following news article:

Context: **In early October 2023**, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.

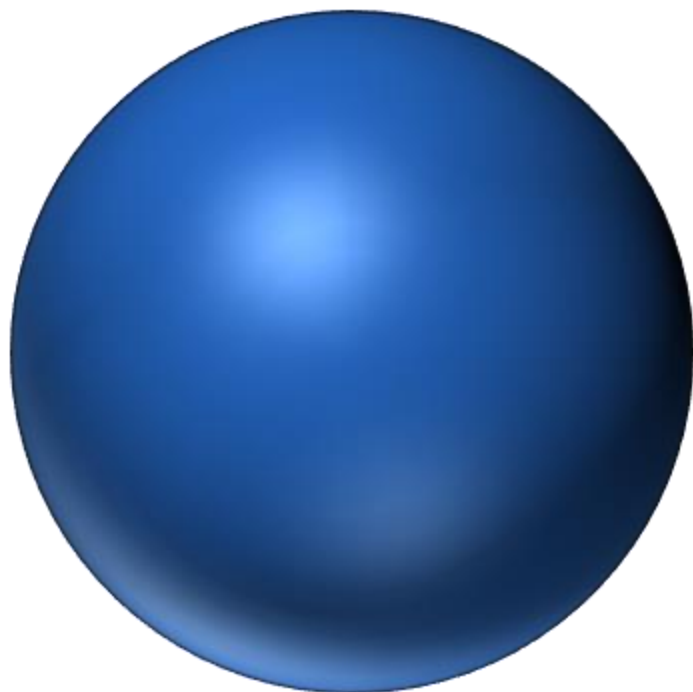


Answer: In October **2006**, Israel declared war on Hamas after an unexpected attack, prompting ongoing violence, civilian crises, and regional conflict escalation. ❌

(b) Faithfulness Hallucination

Traditional LLMs

Model size



Test-time compute



Reasoning LLMs

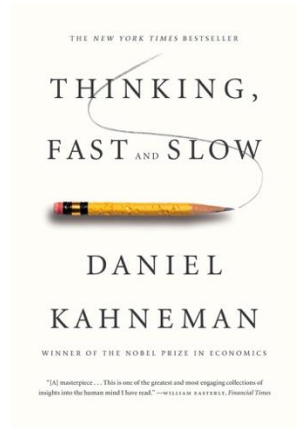
Model size



Test-time compute



Towards utilizing Test-Time compute



SYSTEM 1
Intuition & instinct

SYSTEM 2
Rational thinking

95%

Unconscious
Fast
Associative
Automatic pilot



5%

Takes effort
Slow
Logical
Lazy
Indecisive

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Chain-of-Thought Prompting (CoT)

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

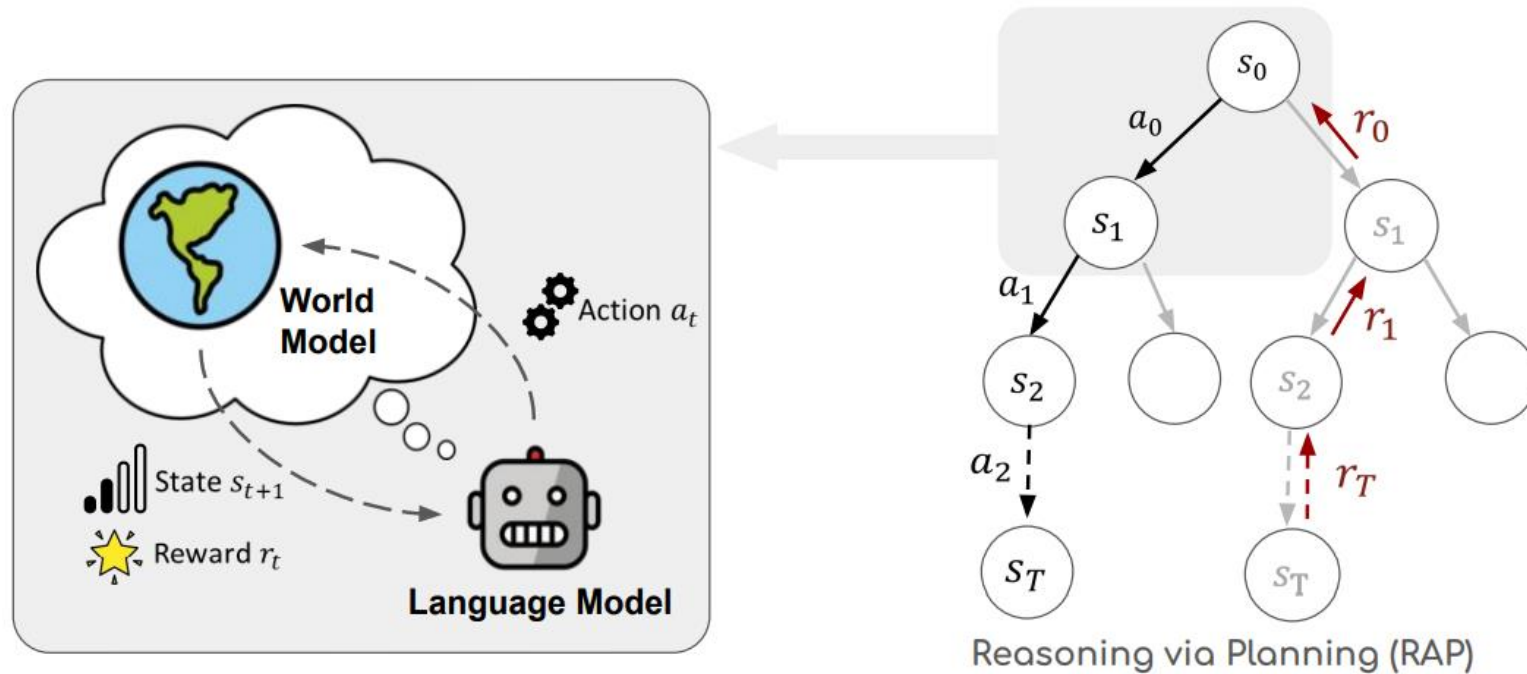
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

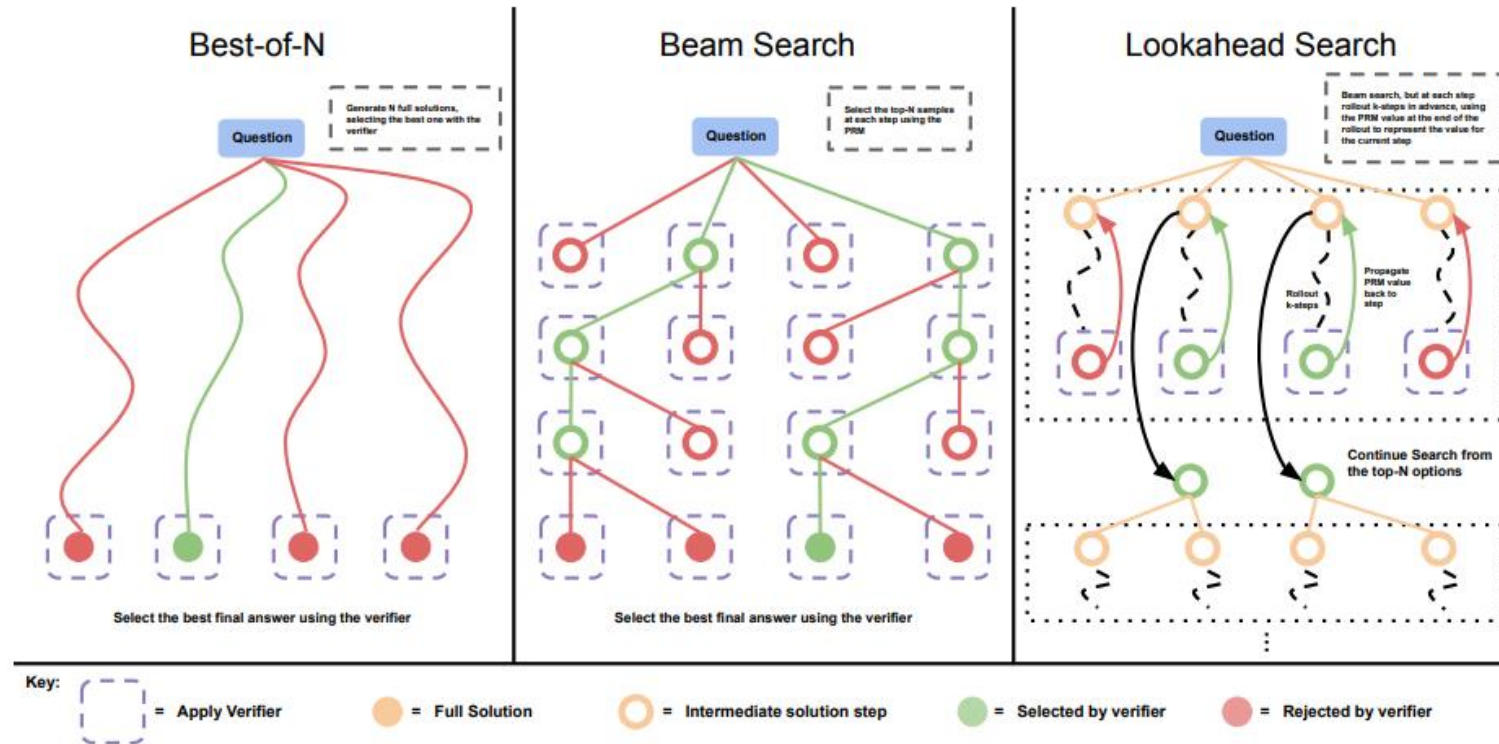
[7] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (arXiv:2201.11903). arXiv. <https://doi.org/10.48550/arXiv.2201.11903>

Reasoning with Language Model is Planning with World Model



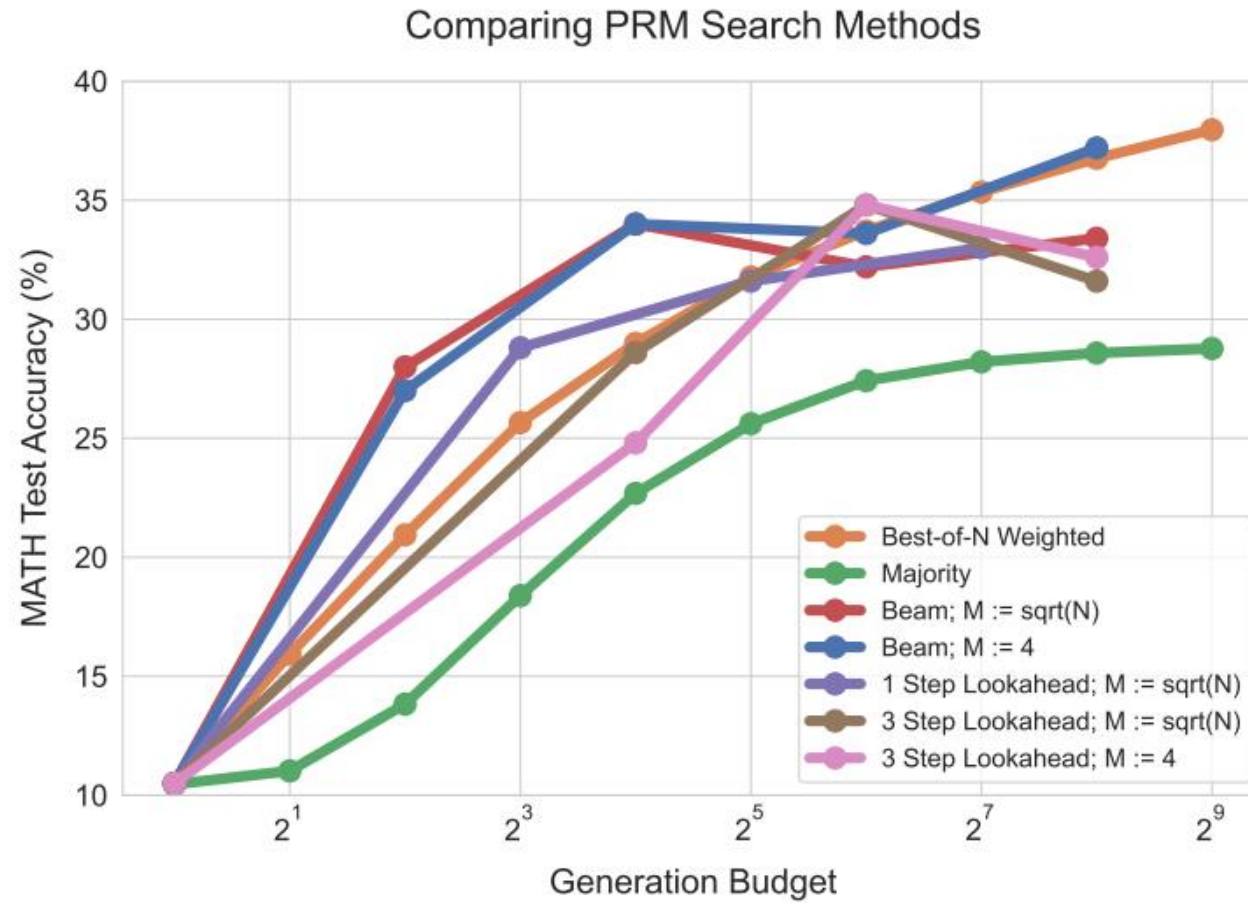
[3] Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., & Hu, Z. (2023). *Reasoning with Language Model is Planning with World Model* (arXiv:2305.14992). arXiv. <https://doi.org/10.48550/arXiv.2305.14992>

Search-based methods



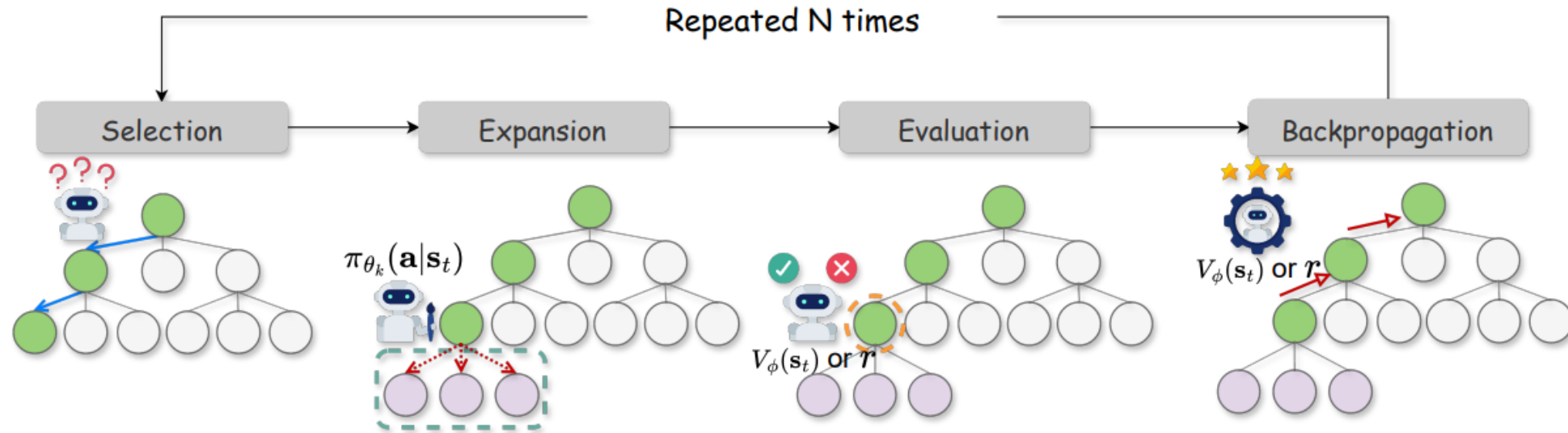
Outcome-supervised reward models (ORM) vs Process-supervised reward models (PRM)

[2] Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). *Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters* (arXiv:2408.03314). arXiv. <https://doi.org/10.48550/arXiv.2408.03314>



[2] Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). *Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters* (arXiv:2408.03314). arXiv. <https://doi.org/10.48550/arXiv.2408.03314>

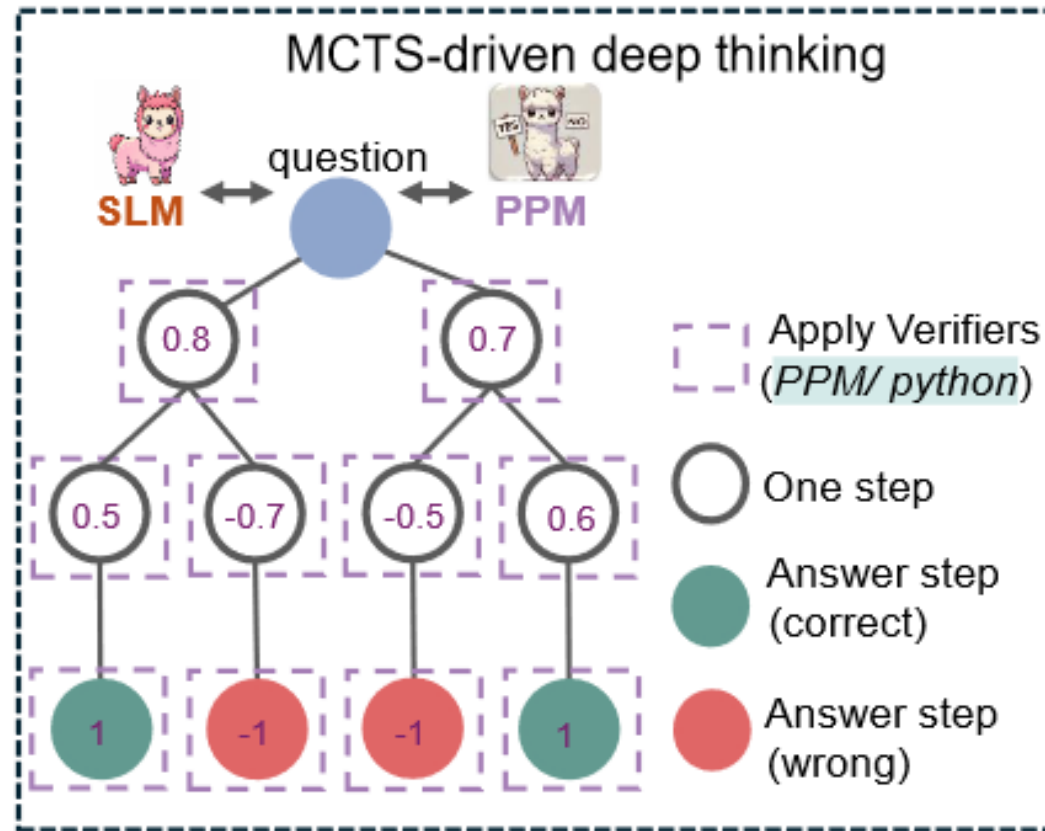
Monte-Carlo Tree Search (MCTS)



$$\text{UCT}(s) = Q(s) + c \sqrt{\frac{\ln N_{\text{parent}}(s)}{N(s)}}; \quad \text{where} \quad Q(s) = \frac{q(s)}{N(s)}$$

[4] Chen, G., Liao, M., Li, C., & Fan, K. (2024). *AlphaMath Almost Zero: Process Supervision without Process* (arXiv:2405.03553). arXiv.
<https://doi.org/10.48550/arXiv.2405.03553>

rStar-Math



[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Problem-solving with application of the Vieta's formulas and the AM-GM inequality

Question: Let a be a positive real number such that all the roots of

$$x^3 + ax^2 + ax + 1 = 0$$

are real. Find the smallest possible value of a .

Our trajectory for proving:

<code>

Step 1: Define the polynomial equation
from sympy import symbols, solve, Eq

Define the variable and the parameter
 $x, a = \text{symbols}('x a')$
<end_of_step> [PPM Score: -0.46430]

Step 2: Define the polynomial equation
 $\text{polynomial} = x**3 + a*x**2 + a*x + 1$
<end_of_step> [PPM Score: 0.53937]

Step 3: Use Vieta's formulas to relate the coefficients to the roots

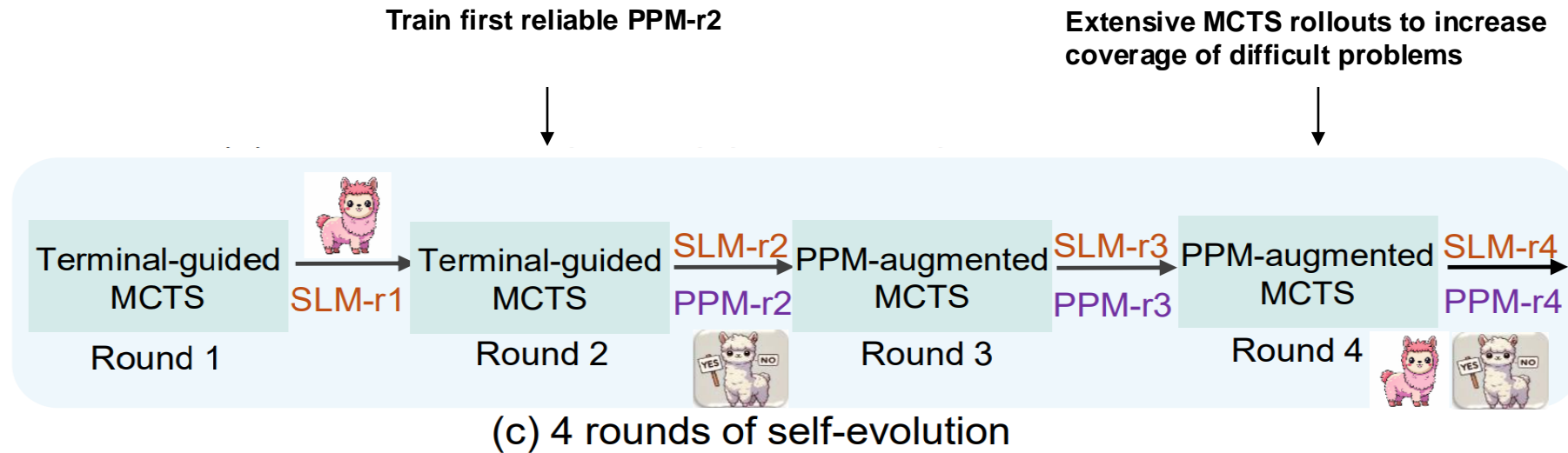
Since the polynomial is monic (leading coefficient is 1), the sum of the roots (taken one at a time) is $-a$
The sum of the products of the roots taken two at a time is a
The product of the roots is -1
Since all roots are real, we can denote them as r_1, r_2, r_3
From Vieta's formulas:
$r_1 + r_2 + r_3 = -a$
$r_1*r_2 + r_2*r_3 + r_3*r_1 = a$
$r_1*r_2*r_3 = -1$
<end_of_step> [PPM Score: 0.9989]

Step 4: Use the AM-GM inequality to find the minimum value of a

By AM-GM, $(r_1*r_2 + r_2*r_3 + r_3*r_1) / 3 \geq (r_1*r_2*r_3)^{(2/3)}$
Substituting the known values:
$a / 3 \geq (-1)^{(2/3)}$
$a / 3 \geq 1$
$a \geq 3$
<end_of_step> [PPM Score: 0.9999]

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Self-Evolved Deep Thinking



DeepSeek-Coder-V2-Instruct (236B)
used to bootstrap SFT data


PPM-augmented MCTS significantly
improves SFT data quality

Improved SFT data quality and increased train set coverage after each round

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Step-by-step verified reasoning trajectories

Question: Bill walks $\frac{1}{2}$ mile south, then $\frac{3}{4}$ mile east, and finally $\frac{1}{2}$ mile south. How many miles is he, in a direct line, from his starting point? Express your answer as a decimal to the nearest hundredth.

Step 1: Calculate the total distance walked south |  NL CoT as Python Comment

```
total_south = 1/2 + 1/2
```

```
# Step 2: Calculate the total distance walked east
```

```
total_east = 3/4
```

```
# Step 3: Use the Pythagorean theorem to find the direct distance from the starting point
```

```
import math
```

```
direct_distance = math.sqrt(total_south**2 + total_east**2)
```

```
# Step 4: Round the direct distance to the nearest hundredth
```

```
direct_distance_rounded = round(direct_distance, 2)
```

```
From the result, we can see that the direct distance from the starting point is  $\boxed{1.25}$  miles
```

Python code execution for step 1:

```
# Step 1: Calculate the total distance walked south
```

```
total_south = 1/2 + 1/2
```

Python code execution for step 2:

```
# Step 1: Calculate the total distance walked south
```

```
total_south = 1/2 + 1/2
```

```
# Step 2: Calculate the total distance walked east
```

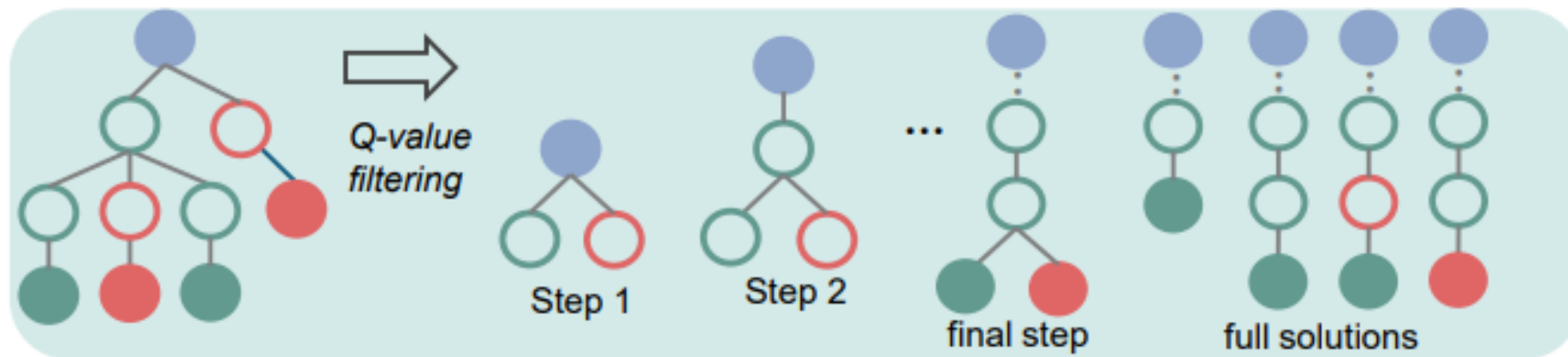
```
total_east = 3/4
```

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Process Preference Model (PPM)

Pairwise ranking loss

$$\mathcal{L}_{ppm}(\theta) = -\frac{1}{2 \times 2} E_{(x, y_i^{pos}, y_i^{neg} \in \mathbb{D})} [\log(\sigma(r_\theta(x, y_i^{pos}) - r_\theta(x, y_i^{neg})))]$$



(b) Construction of per-step preference pairs based on Q-values

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Benchmark results

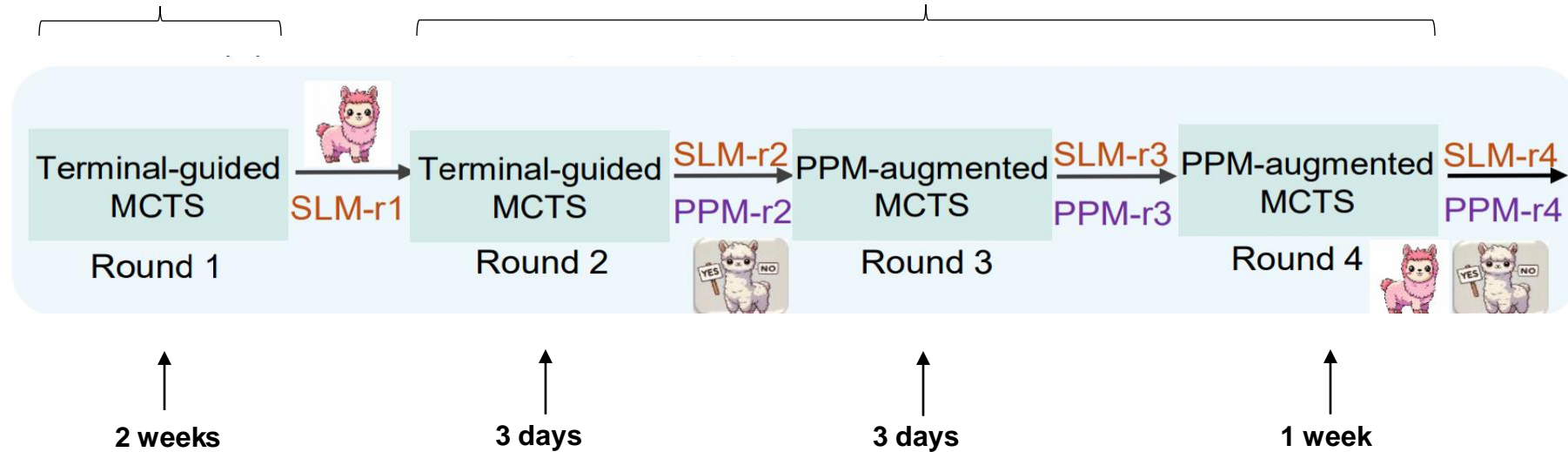
Task (pass@1 Acc)	rStar-Math (Qwen-7B)	rStar-Math (Qwen-1.5B)	rStar-Math (Phi3-mini)	OpenAI o1-preview	OpenAI o1-mini	QWQ 32B-preview	GPT-4o	DeepSeek-V3
MATH	90.0	88.6	86.4	85.5	90.0	90.6	76.6	90.2
AIME 2024	53.3	46.7	43.3	44.6	56.7	50.0	9.3	39.2
Olympiad Bench	65.6	64.6	60.3	-	65.3	61.2	43.3	55.4
College Math	60.5	59.3	59.1	-	57.8	55.8	48.5	58.9
Omni-Math	50.5	48.5	46.0	52.5	60.5	49.6	30.5	35.9

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Training cost

10 nodes of 8x80GB H100 GPUs

15 nodes of 4x40GB A100 GPUs



[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Training cost

Amount	GPU model	Power (per GPU)	Total power	Energy used	Estimated electricity cost (\$0.20/kWh)	Cloud cost (per GPU)	Estimated total cloud cost
80x	80GB H100	700 W	56 kW	18,816 kWh	~\$3,800	~ \$1.90/hr - \$3.50/hr	~ \$51,072 - \$94,080
60x	40GB A100	400 W	24 kW	7,488 kWh	~\$1,500	~ \$1.00/hr - \$1.80/hr	~ \$18,720 - \$33,696

Datasheets:

H100: <https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet>

A100: <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet.pdf>

Cloud costs:

H100: <https://getdeploying.com/reference/cloud-gpu/nvidia-h100>

A100: <https://getdeploying.com/reference/cloud-gpu/nvidia-a100>

Inference cost

Table 9: Inference costs of rStar-Math. We show the average number of generated tokens required to generate a trajectory for a given question.

MATH	AIME 2024	AMC 2023	Olympiad Bench	College Math	GSM8K	GaokaoEn 2023
5453	15693	14544	7889	4503	3299	6375

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Question: Given positive integers x and y such that $2x^2y^3 + 4y^3 = 149 + 3x^2$, what is the value of $x + y$?

Low-quality Steps

```
<code>
from sympy import symbols, Eq, solve
# Define the variables x and y
x, y = symbols('x y')
<end_of_step>
```

PPM score: -0.08

```
# Define the equation
equation = Eq(2*x**2*y**3 + 4*y**3, 149 + 3*x**2)
<end_of_step>
```

PPM score: -0.219

```
# Solve the equation for y in terms of x
solution = solve(equation, y)
<end_of_step>
```

PPM score: -0.348

```
# Now print the final answer
print(solution)\n
<end_of_code>
<output>
[(3*x**2/(2*x**2 + 4) + 149/(2*x**2 + 4))**(1/3), -(3*x**2/(2*x**2 + 4) + 149/(2*x**2 + 4))**(1/3)/2 - sqrt(3)*I*(3*x**2/(2*x**2 + 4) + 149/(2*x**2 + 4))**(1/3)/2, -(3*x**2/(2*x**2 + 4) + 149/(2*x**2 + 4))**(1/3)/2 + sqrt(3)*I*(3*x**2/(2*x**2 + 4) + 149/(2*x**2 + 4))**(1/3)/2]
<end_of_output>
```

Wrong PPM score: -0.517

```
<answer>
The value of  $\|(x + y)\|$  is  $\boxed{8}$ .
<end_of_answer>
```

Wrong PPM score: -0.529

Intrinsic self-reflection:
Thinking outside the box,
find an easier solution!

```
# Since we need positive integers, we need to find the values of x and y that satisfy the equation
# We will check for small integer values of x and y
for x_val in range(1, 10):
    for y_val in range(1, 10):
        if 2*x_val**2*y_val**3 + 4*y_val**3 == 149 + 3*x_val**2:
            print(f"x = {x_val}, y = {y_val}")
            print(f"x + y = {x_val + y_val}")
            break
<end_of_step>
```

PPM score: 0.620

```
# Now print the final answer
# The loop will find the first valid pair (x, y) and print x + y
<end_of_code>
<output>
x = 3, y = 2
x + y = 5
<end_of_output>
```

PPM score: 0.698

```
<answer>
From the result, we can see that the value of x + y is  $\boxed{5}$ 
<end_of_answer>
```

Correct PPM score: 0.835

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

References

- [1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>
- [2] Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). *Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters* (arXiv:2408.03314). arXiv. <https://doi.org/10.48550/arXiv.2408.03314>
- [3] Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., & Hu, Z. (2023). *Reasoning with Language Model is Planning with World Model* (arXiv:2305.14992). arXiv. <https://doi.org/10.48550/arXiv.2305.14992>
- [4] Chen, G., Liao, M., Li, C., & Fan, K. (2024). *AlphaMath Almost Zero: Process Supervision without Process* (arXiv:2405.03553). arXiv. <https://doi.org/10.48550/arXiv.2405.03553>
- [5] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023). *Let's Verify Step by Step* (arXiv:2305.20050). arXiv. <https://doi.org/10.48550/arXiv.2305.20050>
- [6] Qi, Z., Ma, M., Xu, J., Zhang, L. L., Yang, F., & Yang, M. (2024). *Mutual Reasoning Makes Smaller LLMs Stronger Problem-Solvers* (arXiv:2408.06195). arXiv. <https://doi.org/10.48550/arXiv.2408.06195>
- [7] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (arXiv:2201.11903). arXiv. <https://doi.org/10.48550/arXiv.2201.11903>
- [8] Zhang, D., Zhou, S., Hu, Z., Yue, Y., Dong, Y., & Tang, J. (2024). *ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search* (arXiv:2406.03816). arXiv. <https://doi.org/10.48550/arXiv.2406.03816>
- [9] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models* (arXiv:2305.10601). arXiv. <https://doi.org/10.48550/arXiv.2305.10601>

Benchmarks

Easy

Hard

MATH	OlympiadBench	AIME 2024
Let $p(x)$ be a cubic polynomial such that $p(2)=0$, $p(-1)=0$, $p(4)=6$, and $p(5)=8$. Find $p(7)$.	Find all triples of (x,y,z) of positive integers such that $x \leq y \leq z$ and $x^3(y^3+z^3)=2012(xyz+2)$	Quadratic polynomials $P(x)$ and $Q(x)$ have leading coefficients 2 and -2 , respectively. The graphs of both polynomials pass through the two points $(16,54)$ and $(20,53)$. Find $P(0) + Q(0)$.
A 6-sided die is weighted so that the probability of any number being rolled is proportional to the value of the roll. (So, for example, the probability of a 2 being rolled is twice that of a 1 being rolled.) What is the expected value of a roll of this weighted die? Express your answer as a common fraction.	Given a positive integer n , determine the largest real number μ satisfying the following condition: for every n -point configuration C in an open unit square U , there exists an open rectangle in U , whose sides are parallel to those of U , which contains exactly one point of C , and has an area greater than or equal to μ .	A circle with radius 6 is externally tangent to a circle with radius 24 . Find the area of the triangular region bounded by the three common tangent lines of these two circles.
The lengths of two opposite sides of a square are decreased by 40% while the lengths of the other two sides are increased by 50% to form a rectangle. By what percent does the square's area decrease?	A circle ω of radius 1 is given. A collection T of triangles is called good, if the following conditions hold: (i) each triangle from T is inscribed in ω ; (ii) no two triangles from T have a common interior point. Determine all positive real numbers t such that, for each positive integer n , there exists a good collection of n triangles, each of perimeter greater than t .	A straight river that is 264 meters wide flows from west to east at a rate of 14 meters per minute. Melanie and Sherry sit on the south bank of the river with Melanie a distance of D meters downstream from Sherry. Relative to the water, Melanie swims at 80 meters per minute, and Sherry swims at 60 meters per minute. At the same time, Melanie and Sherry begin swimming in straight lines to a point on the north bank of the river that is equidistant from their starting positions. The two women arrive at this point simultaneously. Find D .

MATH: URL <https://huggingface.co/datasets/HuggingFaceH4/MATH-500>
 OlympiadBench: URL <https://github.com/OpenBMB/OlympiadBench?tab=readme-ov-file>
 AIME 2024: URL <https://huggingface.co/datasets/AI-MO/aimo-validation-aime>.

Ablation study – Self-Evolved Deep Thinking

Table 6: The continuously improved math reasoning capabilities through rStar-Math self-evolved deep thinking. Starting from round 2, the 7B base model powered by rStar-Math surpasses GPT-4o.

Round#	MATH	AIME 2024	AMC 2023	Olympiad Bench	College Math	GSM8K	GaokaoEn 2023
GPT-4o	76.6	9.3	47.5	43.3	48.5	92.9	67.5
Base 7B model	58.8	0.0	22.5	21.8	41.6	91.6	51.7
rStar-Math Round 1	75.2	10.0	57.5	35.7	45.4	90.9	60.3
rStar-Math Round 2	86.6	43.3	75.0	59.4	55.6	94.0	76.4
rStar-Math Round 3	87.0	46.7	80.0	61.6	56.5	94.2	77.1
rStar-Math Round 4	89.4	50.0	87.5	65.3	59.0	95.0	80.5

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Ablation study – Step-by-step verified reasoning trajectories

Table 7: Ablation study on the effectiveness of our step-by-step verified reasoning trajectories as the SFT dataset. We report the SFT accuracy of Qwen2.5-Math-7B fine-tuned with different datasets.

	Dataset	MATH	AIME	AMC	Olympiad Bench	College Math	GSM8K	GaokaoEn 2023
GPT-4o	-	76.6	9.3	47.5	43.3	48.5	92.9	67.5
GPT4-distillation (Open-sourced)	MetaMath	55.2	3.33	32.5	19.1	39.2	85.1	43.6
	NuminaMath-CoT	69.6	10.0	50.0	37.2	43.4	89.8	59.5
Self-generation by policy SLM-r3	Random sample	72.4	10.0	45.0	41.0	48.0	87.5	57.1
	Rejection sampling	73.4	13.3	47.5	44.7	50.8	89.3	61.7
	Step-by-step verified (ours)	78.4	26.7	47.5	47.1	52.5	89.7	65.7

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Ablation study – Process Preference Model (PPM)

Table 8: Ablation study on the reward model. Process reward models (PQM and PPM) outperform ORM, with PPM pushing the frontier of math reasoning capabilities.

RM	Inference	MATH	AIME	AMC	Olympiad Bench	College Math	GSM8K	GaokaoEn
o1-mini	-	90.0	56.7	95.0	65.3	55.6	94.8	78.6
ORM	Best-of-N	82.6	26.7	65.0	55.1	55.5	92.3	72.5
PQM	MCTS	88.2	46.7	85.0	62.9	57.6	94.6	79.5
PPM	MCTS	89.4	50.0	87.5	65.3	59.0	95.0	80.5

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Surpassing bootstrap model

Table 3: Pass@1 accuracy of the resulting policy SLM in each round, showing continuous improvement until surpassing the bootstrap model.

Round#	MATH	AIME 2024	AMC 2023	Olympiad Bench	College Math	GSM8K	GaokaoEn 2023
DeepSeek-Coder-V2-Instruct (bootstrap model)	75.3	13.3	57.5	37.6	46.2	94.9	64.7
Base (Qwen2.5-Math-7B)	58.8	0.0	22.5	21.8	41.6	91.6	51.7
policy SLM-r1	69.6	3.3	30.0	34.7	44.5	88.4	57.4
policy SLM-r2	73.6	10.0	35.0	39.0	45.7	89.1	59.7
policy SLM-r3	75.8	16.7	45.0	44.1	49.6	89.3	62.8
policy SLM-r4	78.4	26.7	47.5	47.1	52.5	89.7	65.7

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Scaling Test-Time Compute

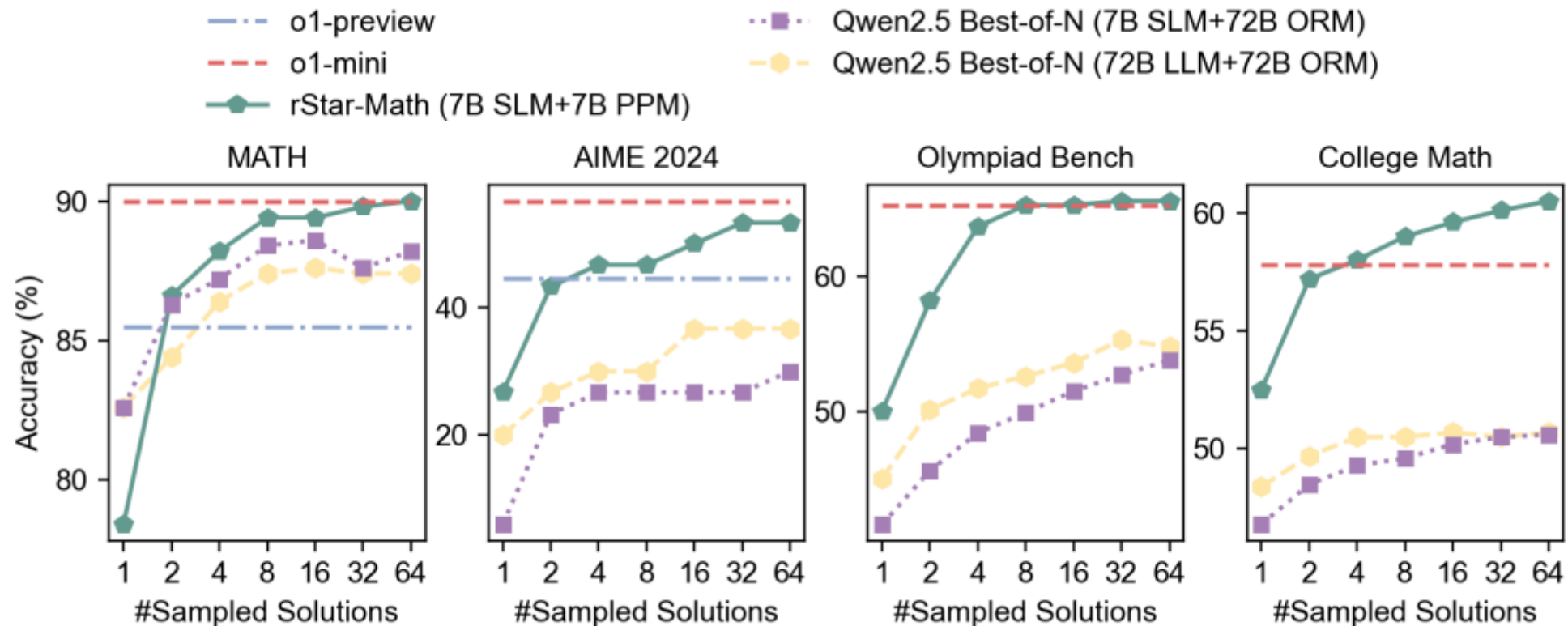


Figure 3: Reasoning performance under scaling up the test-time compute.

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

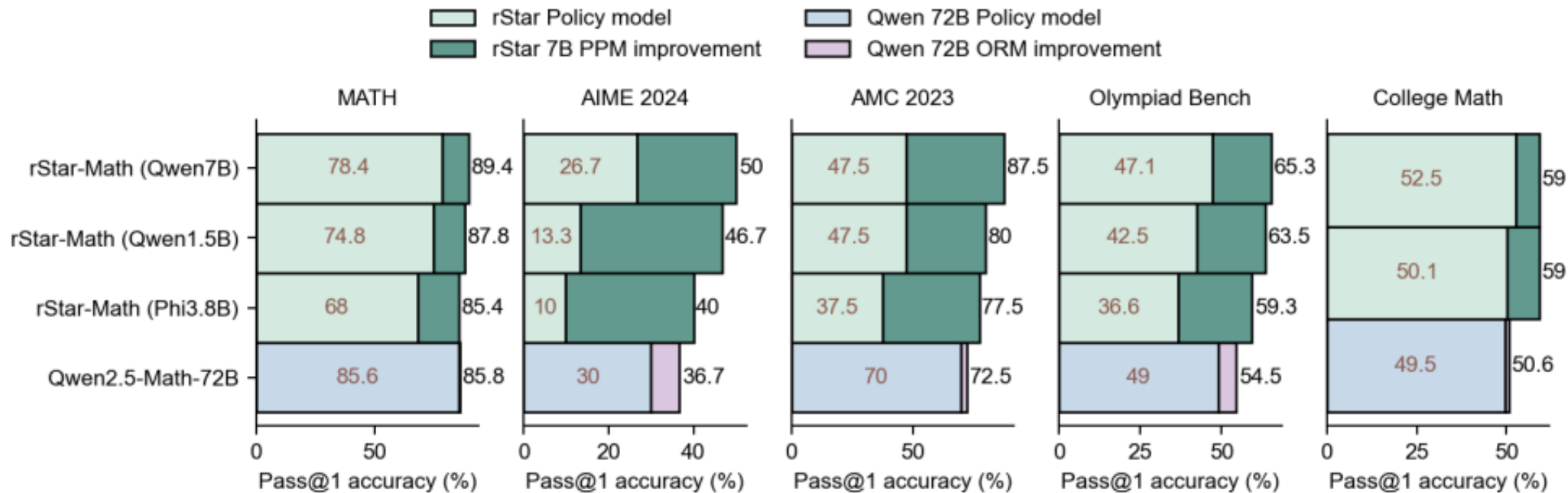


Figure 5: Pass@1 accuracy of policy models and their accuracy after applying System 2 reasoning with various reward models, shows that reward models primarily determine the final performance.

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Table 10: Pass@1 (greedy) accuracy of our fine-tuned policy models for Phi3-mini, Qwen2.5-Math-1.5B, Qwen2-Math-7B and Qwen2.5-Math-7B.

Model	MATH	AIME 2024	AMC 2023	Olympiad Bench	College Math	GSM8K	GaokaoEn 2023
<i>General Base Model: Phi3-mini-Instruct (3.8B)</i>							
Phi3-mini-Instruct	41.4	3.33	7.5	12.3	33.1	85.7	37.1
Our policy model	68.0	10.0	37.5	36.6	48.7	87.9	53.2
<i>Math-Specialized Base Model: Qwen2.5-Math-1.5B</i>							
Qwen2.5-Math-1.5B	51.2	0.0	22.5	16.7	38.4	74.6	46.5
Qwen2.5-Math-1.5B-Instruct	60.0	10.0	60.0	38.1	47.7	84.8	65.5
Our policy model	74.8	13.3	47.5	42.5	50.1	83.1	58.7
<i>Math-Specialized Base Model: Qwen2-Math-7B</i>							
Qwen2-Math-7B	53.4	3.3	25.0	17.3	39.4	80.4	47.3
Qwen2-Math-7B-Instruct	73.2	13.3	62.5	38.2	45.9	89.9	62.1
Our policy model	73.8	16.7	45.0	43.9	52.0	88.3	65.2
<i>Math-Specialized Base Model: Qwen2.5-Math-7B</i>							
Qwen2.5-Math-7B	58.8	0.0	22.5	21.8	41.6	91.6	51.7
Qwen2.5-Math-7B-Instruct	82.6	6.0	62.5	41.6	46.8	95.2	66.8
Our policy model	78.4	26.7	47.5	47.1	52.5	89.7	65.7

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>

Table 2: Percentage of the 747k math problems correctly solved in each round. Only problems have correct solutions are included in the training set. The first round uses DeepSeek-Coder-Instruct as the policy LLM, while later rounds use our fine-tuned 7B policy SLM.

#	models in MCTS	GSM-level	MATH-level	Olympiad-level	All
Round 1	DeepSeek-Coder-V2-Instruct	96.61%	67.36%	20.99%	60.17%
Round 2	policy SLM-r1	97.88%	67.40%	56.04%	66.60%
Round 3	policy SLM-r2, PPM-r2	98.15%	88.69%	62.16%	77.86%
Round 4	policy SLM-r3, PPM-r3	98.15%	94.53%	80.58%	90.25%

[1] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. (2025). *rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking* (arXiv:2501.04519). arXiv. <https://doi.org/10.48550/arXiv.2501.04519>