

SimpleRL: On Emerging Reasoning of Open Models in the Wild

Weihao Zeng^{*}, Yuzhen Huang^{*}, Qian Liu^{*}, Wei Liu, Keqing He, Zejun Ma, Junxian He

The Hong Kong University of Science and Technology

C) Star 3,509





Long Chain of Thought (CoT)

O1 Solution

Okay, let's see. So Aya goes for a 9-kilometer walk every morning and stops at a coffee shop. The problem gives two scenarios with different speeds and total times, and we need to find the total time when she walks at a different speed. Hmm.

First, I need to parse the information given. The key points are:

1. The distance is always 9 km.

2. The total time includes both walking time and t minutes at the coffee shop.

3. Two different speeds are given, each resulting in different total times, but the coffee shop time t is the same in both cases.

Wait, let me make sure. The problem states when she walks at s km/h, the total time is 4 hours (which includes t minutes). When she walks at s + 2 km/h, the total time is 2 hours and 24 minutes (which also includes t minutes). So both cases have the same t, but different walking times leading to different total times.

So, the total time is walking time plus t. Since t is in minutes, but the total time is given in hours and minutes, I need to convert everything to the same units. Maybe convert all times to minutes to avoid confusion.

.....

Since speed can't be negative, s = (5)/2 = 2.5 km/h

Wait, sqrt(49)=7, so (-2 +7)/2=5/2=2.5, yes. So s=2.5 km/h.

Now check if that makes sense.

.....

Third scenario: speed is s +0.5=3 km/h. Walking time=9/3=3 hours=180 minutes. Add t=24, total=204.

Yes, seems correct.

The Hong Kong University of Science and Technology

2



Long Chain of Thought (CoT)

O1 Solution

Okay, let's see. So Aya goes for a 9-kilometer walk every morning and stops at a coffee shop. The problem gives two scenarios with different speeds and total times, and we need to find the total time when she walks at a different speed. Hmm.

First, I need to parse the information given. The key points are:

1. The distance is always 9 km.

2. The total time includes both walking time and t minutes at the coffee shop.

3. Two different speeds are given, each resulting in different total times, but the coffee shop time t is the same in both cases.

Wait, let me make sure. The problem states when she walks at s km/h, the total time is 4 hours (which includes t minutes). When she walks at s + 2 km/h, the total time is 2 hours and 24 minutes (which also includes t minutes). So both cases have the same t, but different walking times leading to different total times.

So, the total time is walking time plus t. Since t is in minutes, but the total time is given in hours and minutes, I need to convert everything to the same units. Maybe convert all times to minutes to avoid confusion.

.....

Since speed can't be negative, s = (5)/2 = 2.5 km/h

Wait, sqrt(49)=7, so (-2 +7)/2=5/2=2.5, yes. So s=2.5 km/h.

Now check if that makes sense.

.....

Third scenario: speed is s +0.5=3 km/h. Walking time=9/3=3 hours=180 minutes. Add t=24, total=204.

Yes, seems correct.

The Hong Kong University of Science and Technology

Long Chain of Thought (CoT)

+ certain cognitive behaviors (e.g., Self-Reflection)

2



DeepSeek-R1 and Kimi-k1.5 choose the extremely simple recipe — Reinforcement Learning



DeepSeek-R1 and Kimi-k1.5 choose the extremely simple recipe — Reinforcement Learning





DeepSeek-R1 and Kimi-k1.5 choose the extremely simple recipe — Reinforcement Learning





DeepSeek-R1 and Kimi-k1.5 choose the extremely simple recipe — Reinforcement Learning



Model naturally develops long CoT during RL





RL Training is not a new thing



RL Training is not a new thing

But the emergence of Long CoT plus Self Reflection is new



Deepseek-R1 conducts experiments on huge models in a large-scale RL setting. What about small model and limited data?





Deepseek-R1 conducts experiments on huge models in a large-scale RL setting. What about small model and limited data?

Does RL works for small model and limited data?

Does long CoT and self-reflection emerge in this case?





RL is applied directly to the base model using 8K MATH (query, answer) pairs without SFT.



RL is applied directly to the base model using 8K MATH (query, answer) pairs without SFT.



[1] Zeng et al. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient. 2025

Qwen2.5-Math-7B-Base Qwen-2.5-Math-7B-Instruct Eurus-2-7B-PRIME Qwen2.5-7B-SimpleRL-Zero

Qwen2.5-Math-7B-Base-SFT

Avg.

Minerva Math OlympiadBench



RL is applied directly to the base model using 8K MATH (query, answer) pairs without SFT.





RL is applied directly to the base model using 8K MATH (query, answer) pairs without SFT.





RL is applied directly to the base model using 8K MATH (query, answer) pairs without SFT.





RL is applied directly to the base model using 8K MATH (query, answer) pairs without SFT.





RL is applied directly to the base model using 8K MATH (query, answer) pairs without SFT.





RL is applied directly to the base model using 8K MATH (query, answer) pairs without SFT.





RL is applied directly to the base model using 8K MATH (query, answer) pairs without SFT.



Achieve gains of nearly 20 absolute points from base model.

Outperform Instruct model and achieves comparable results to PRIME with 50x data efficiency.



• Start from the Qwen2.5-Math-7B-Base model.

	Qwen2.5-Math- 7B-Instruct	rStar-Math-7B	Eurus-2-7B-PRIME	Qwen2.5-7B- SimpleRL-Zero
Base Model	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B
SFT Data	2.5M (open- source and in- house)	~7.3 M (MATH, NuminaMath, etc.)	230K	0
RM Data	618K (in-house)	~7k (in-house)	0	0
RM	Qwen2.5-Math- RM (72B)	None	Eurus-2-7B-SFT	None
RL Data	66K queries × 32 samples	~3.647 M × 16	150K queries × 4 samples	8K queries × 8 samples



• Start from the Qwen2.5-Math-7B-Base model.

	Qwen2.5-Math- 7B-Instruct	rStar-Math-7B	Eurus-2-7B-PRIME	Qwen2.5-7B- SimpleRL-Zero
Base Model	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B
SFT Data	2.5M (open- source and in- house)	~7.3 M (MATH, NuminaMath, etc.)	230K	0
RM Data	618K (in-house)	~7k (in-house)	0	0
RM	Qwen2.5-Math- RM (72B)	None	Eurus-2-7B-SFT	None
RL Data	66K queries × 32 samples	~3.647 M × 16	150K queries × 4 samples	8K queries × 8 samples

Using only 8K examples from the original MATH dataset, much less than other methods.



• Start from the Qwen2.5-Math-7B-Base model.

	Qwen2.5-Math- 7B-Instruct	rStar-Math-7B	Eurus-2-7B-PRIME	Qwen2.5-7B- SimpleRL-Zero
Base Model	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B
SFT Data	2.5M (open- source and in- house)	~7.3 M (MATH, NuminaMath, etc.)	230K	0
RM Data	618K (in-house)	\sim 7k (in-house)	0	0
RM	Qwen2.5-Math- RM (72B)	None	Eurus-2-7B-SFT	None
RL Data	66K queries × 32 samples	~3.647 M × 16	150K queries × 4 samples	8K queries × 8 samples

Using only 8K examples from the original MATH dataset, much less than other methods.



Start from the Qwen2.5-Math-7B-Base model.

	Qwen2.5-Math- 7B-Instruct	rStar-Math-7B	Eurus-2-7B-PRIME	Qwen2.5-7B- SimpleRL-Zero
Base Model	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B
SFT Data	2.5M (open- source and in- house)	~7.3 M (MATH, NuminaMath, etc.)	230K	0
RM Data	618K (in-house)	~7k (in-house)	0	0
RM	Qwen2.5-Math- RM (72B)	None	Eurus-2-7B-SFT	None
RL Data	66K queries × 32 samples	~3.647 M × 16	150K queries × 4 samples	8K queries × 8 samples

Using only 8K examples from the original MATH dataset, much less than other methods. Evaluate on challenging math benchmarks, AIME, AMC ...















Almost all open reproduction of DeepSeek-R1 is based Qwen models, but Qwen models are kinda special.





Almost all open reproduction of DeepSeek-R1 is based Qwen models, but Qwen models are kinda special.

Are the conclusions from these works Qwen-specific?





Consistent Results



[1] Zeng et al. SimpleRL-Zoo: Investigating and Taming Zero Reinforcement Learning for Open Base Models in the Wild. 2025

Main Results

GSM8K	MATH 500	Minerva Math	Olympiad Bench	AIME24 (Pass@1)	AIME24 (Avg@32)	AMC23	Avg.		
Llama, DeepSeek and Mistral Models									
21.2	4.2	4.0	2.4	0.0	0.0	0.0	5.3		
75.0	15.8	6.6	4.1	0.0	0.2	10.0	18.6		
39.7	13.6	4.8	3.1	0.0	0.2	2.5	10.6		
79.2	23.0	9.6	5.3	0.0	0.2	15.0	22.0		
28.4	19.4	5.5	4.7	0.0	0.0	10.0	11.3		
78.5	39.6	21.0	12.6	3.3	0.6	20.0	29.2		
78.6	43.6	10.7	11.6	3.3	0.5	17.5	27.6		
92.0	70.6	36.8	36.6	16.7	13.1	45.0	49.6		
Owen Series Models									
36.7	15.8	\sim 4.8	2.8	0.0	0.3	12.5	12.1		
49.5	34.4	10.3	8.9	0.0	0.7	22.5	20.9		
55.7	29.6	6.6	6.5	0.0	0.1	12.5	18.5		
74.4	59.0	20.2	21.0	6.7	4.2	35.0	36.1		
88.2	64.6	25.7	30.1	3.3	0.3	30.0	40.3		
91.7	78.2	38.6	40.4	20.0	15.6	62.5	55.2		
65.5	63.6	12.5	25.8	13.3	8.6	42.5	37.2		
90.2	80.2	37.5	39.0	40.0	24.0	70.0	59.5		
91.6	65.4	24.3	33.5	6.7	3.4	37.5	43.2		
94.4	80.2	40.4	44.9	23.3	14.2	57.6	56.8		
92.9	68.6	27.9	31.1	10.0	4.5	45.0	45.9		
95.9	82.4	42.6	46.4	36.7	27.2	67.5	61.9		
	GSM8K 21.2 75.0 39.7 79.2 28.4 78.5 78.6 92.0 36.7 49.5 55.7 74.4 88.2 91.7 65.5 90.2 91.6 92.9 95.9	GSM8KMATH 50021.2Llama,21.24.275.015.839.713.679.223.028.419.478.539.678.643.692.070.636.715.849.534.455.729.674.459.088.264.691.778.265.563.690.280.291.665.494.480.292.968.695.982.4	GSM8KMATH 500Minerva Math21.24.24.021.24.24.075.015.86.639.713.64.879.223.09.628.419.45.578.539.621.078.643.610.792.070.636.836.715.84.849.534.410.355.729.66.674.459.020.288.264.625.791.778.238.665.563.612.590.280.237.591.665.424.394.480.240.492.968.627.995.982.442.6	GSM8K MATH 500 Minerva Math Olympiad Bench 21.2 4.2 4.0 2.4 75.0 15.8 6.6 4.1 39.7 13.6 4.8 3.1 79.2 23.0 9.6 5.3 4.7 78.5 39.6 21.0 12.6 4.7 78.5 39.6 21.0 12.6 4.7 78.5 39.6 21.0 12.6 4.7 78.5 39.6 21.0 12.6 4.7 78.5 39.6 21.0 12.6 4.7 78.5 39.6 21.0 12.6 4.7 78.5 39.6 21.0 12.6 4.7 78.5 39.6 10.7 11.6 4.8 2.8 36.7 15.8 4.8 2.8 4.4 4.9 4.9 4.9 4.9 4.9 4.9 4.1 4.8 4.8 4.8 4.8 4.8 4.8 4.8 4.4 <td>GSM8KMATH 500Minerva MathOlympiad BenchAIME24 (Pass@1)21.2Llama, DeepSeek and Mistral Model21.24.24.00.075.015.86.64.10.039.713.64.83.10.079.223.09.65.30.028.419.45.54.70.078.539.621.012.63.378.643.610.711.63.392.070.636.836.616.736.715.84.82.80.055.729.66.66.50.055.729.66.66.50.055.729.66.66.50.074.459.020.221.06.788.264.625.730.13.391.778.238.640.420.065.563.612.525.813.390.280.237.539.040.091.665.424.333.56.794.480.240.444.923.392.968.627.931.110.095.982.442.646.436.7</td> <td>GSM8KMATH 500Minerva MathOlympiad BenchAIME24 (Pass@1)AIME24 (Avg@32)21.24.24.02.40.00.075.015.86.64.10.00.239.713.64.83.10.00.239.713.64.83.10.00.223.09.65.30.00.228.419.45.54.70.00.078.539.621.012.63.30.678.643.610.711.63.30.592.070.636.836.616.713.157.729.66.66.50.00.155.729.66.66.50.00.175.459.020.221.06.74.236.715.84.82.80.00.155.729.66.66.50.00.175.459.020.221.06.74.288.264.625.730.13.30.391.778.238.640.420.015.665.563.612.525.813.38.690.280.237.539.040.024.091.665.424.333.56.73.494.480.240.444.923.314.292.968.627.931.110.04.595.982.442.6<!--</td--><td>GSM8KMATH 500Minerva MathOlympiad BenchAIME24 (Pass@1)AIME24 (Avg@32)AMC2321.24.24.00.40.00.00.075.015.86.64.10.00.210.039.713.64.83.10.00.22.579.223.09.65.30.00.215.028.419.45.54.70.00.010.078.539.621.012.63.30.620.078.643.610.711.63.30.517.592.070.636.836.616.713.145.078.539.410.38.90.00.312.592.070.66.66.50.012.512.555.729.66.66.50.012.512.549.534.410.38.90.112.512.555.729.66.66.50.012.512.555.729.66.66.50.012.513.030.091.778.238.640.420.015.662.562.565.563.612.555.734.642.530.040.024.070.091.778.238.640.420.015.662.565.563.642.565.565.563.612.555.734.635.567.7</td></td>	GSM8KMATH 500Minerva MathOlympiad BenchAIME24 (Pass@1)21.2Llama, DeepSeek and Mistral Model21.24.24.00.075.015.86.64.10.039.713.64.83.10.079.223.09.65.30.028.419.45.54.70.078.539.621.012.63.378.643.610.711.63.392.070.636.836.616.736.715.84.82.80.055.729.66.66.50.055.729.66.66.50.055.729.66.66.50.074.459.020.221.06.788.264.625.730.13.391.778.238.640.420.065.563.612.525.813.390.280.237.539.040.091.665.424.333.56.794.480.240.444.923.392.968.627.931.110.095.982.442.646.436.7	GSM8KMATH 500Minerva MathOlympiad BenchAIME24 (Pass@1)AIME24 (Avg@32)21.24.24.02.40.00.075.015.86.64.10.00.239.713.64.83.10.00.239.713.64.83.10.00.223.09.65.30.00.228.419.45.54.70.00.078.539.621.012.63.30.678.643.610.711.63.30.592.070.636.836.616.713.157.729.66.66.50.00.155.729.66.66.50.00.175.459.020.221.06.74.236.715.84.82.80.00.155.729.66.66.50.00.175.459.020.221.06.74.288.264.625.730.13.30.391.778.238.640.420.015.665.563.612.525.813.38.690.280.237.539.040.024.091.665.424.333.56.73.494.480.240.444.923.314.292.968.627.931.110.04.595.982.442.6 </td <td>GSM8KMATH 500Minerva MathOlympiad BenchAIME24 (Pass@1)AIME24 (Avg@32)AMC2321.24.24.00.40.00.00.075.015.86.64.10.00.210.039.713.64.83.10.00.22.579.223.09.65.30.00.215.028.419.45.54.70.00.010.078.539.621.012.63.30.620.078.643.610.711.63.30.517.592.070.636.836.616.713.145.078.539.410.38.90.00.312.592.070.66.66.50.012.512.555.729.66.66.50.012.512.549.534.410.38.90.112.512.555.729.66.66.50.012.512.555.729.66.66.50.012.513.030.091.778.238.640.420.015.662.562.565.563.612.555.734.642.530.040.024.070.091.778.238.640.420.015.662.565.563.642.565.565.563.612.555.734.635.567.7</td>	GSM8KMATH 500Minerva MathOlympiad BenchAIME24 (Pass@1)AIME24 (Avg@32)AMC2321.24.24.00.40.00.00.075.015.86.64.10.00.210.039.713.64.83.10.00.22.579.223.09.65.30.00.215.028.419.45.54.70.00.010.078.539.621.012.63.30.620.078.643.610.711.63.30.517.592.070.636.836.616.713.145.078.539.410.38.90.00.312.592.070.66.66.50.012.512.555.729.66.66.50.012.512.549.534.410.38.90.112.512.555.729.66.66.50.012.512.555.729.66.66.50.012.513.030.091.778.238.640.420.015.662.562.565.563.612.555.734.642.530.040.024.070.091.778.238.640.420.015.662.565.563.642.565.565.563.612.555.734.635.567.7		

RL Generalizes

Model	IFEVAL strict-prompt	MMLU Stem	MMLU	GPQA	Avg.			
Llama, DeepSeek and Mistral Models								
Mistral-v0.1-7B	13.5	26.1	28.0	23.2	22.7			
\hookrightarrow + SimpleRL-Zoo	21.8	28.1	34.6	30.3	28.7			
Llama-3.1-8B	16.1	27.1	28.7	22.7	23.6			
\hookrightarrow + SimpleRL-Zoo	25.1	40.7	44.5	20.2	32.6			
DeepSeek-Math-7B	11.5	21.6	22.7	19.2	18.7			
\hookrightarrow + SimpleRL-Zoo	16.3	47.4	45.5	27.3	34.1			
Mistral-Small-24B	17.4	30.9	31.7	20.2	25.0			
\hookrightarrow + SimpleRL-Zoo	23.5	73.9	78.8	45.0	55.3			
Owen Series Models								
Qwen-2.5-0.5B	9.6	23.2	24.9	24.8	20.6			
\hookrightarrow + SimpleRL-Zoo	14.4	32.1	34.6	26.3	26.8			
Qwen-2.5-1.5B	15.2	33.1	35.4	24.8	27.1			
\hookrightarrow + SimpleRL-Zoo	20.3	42.1	45.2	28.8	34.1			
Qwen-2.5-7B	21.3	39.8	38.6	23.7	30.8			
\hookrightarrow + SimpleRL-Zoo	25.9	49.6	47.0	29.8	38.1			
Qwen-2.5-Math-7B	14.1	40.6	38.0	27.8	30.1			
\hookrightarrow + SimpleRL-Zoo	17.0	55.6	56.6	35.4	41.1			
Qwen-2.5-14B	22.9	59.8	63.5	24.8	42.7			
\hookrightarrow + SimpleRL-Zoo	29.4	76.3	79.1	50.0	58.7			
Qwen-2.5-32B	24.6	60.7	62.7	38.9	46.7			
\hookrightarrow + SimpleRL-Zoo	31.2	79.0	82.5	49.5	60.6			

Response length increase may be due to repetition and the response does not stop

Response length increase may be due to repetition and the response does not stop

Response length increase may be due to repetition and the response does not stop

Response length increase of Mistral-7B is unhealthy

Cognitive Behavior

[1] Gandhi et al. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. 2025

14

Cognitive Behavior

Response length is just a superficial metric, it does not directly reflect "aha moment". We should focus more on the true metric — We use GPT-40 to judge the cognitive behaviors [1]

[1] Gandhi et al. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. 2025

14

Format Reward is Not Always a Good Thing

Format Reward is Not Always a Good Thing

Applying format reward limits exploration in the initial stage and hurts particularly for weaker base models

Data Difficulty, not too hard, not too easy

(a) Mistral-7b-v0.1

Revisit the pretraining -> short CoT SFT -> RL recipe

Revisit the pretraining -> short CoT SFT -> RL recipe Why RL fails in the past?

Why RL fails in the past? Revisit the pretraining -> short CoT SFT -> RL recipe

Mistral-Small-24B, NuminaMath CoT SFT -> RL

Why RL fails in the past? Revisit the pretraining -> short CoT SFT -> RL recipe

The worse it gets when we perform more SFT before RL

Mistral-Small-24B, NuminaMath CoT SFT -> RL

17

Revisit the pretraining -> short CoT SFT -> RL recipe Why RL fails in the past?

Init Model	GSM8K	MATH 500	Minerva Math	Olympiad Bench	AIME24 (pass@1)	AMC23	Avg.
0 Step	92.0	70.6	36.8	36.6	16.7	45.0	49.6
10 Step	93.0	69.4	39.7	32.3	10.4	44.1	48.2
20 Step	92.6	65.2	34.2	30.7	6.7	38.4	44.6
200 Step	90.3	59.0	31.6	23.3	2.1	26.9	38.9
1000 Step	88.9	48.8	27.6	20.7	2.5	18.1	34.4
2000 Step	89.8	49.0	23.2	18.1	0.8	20.3	33.5
4000 Step	87.7	52.0	23.5	17.2	2.1	21.6	34.0

Mistral-Small-24B, OpenHermes-2.5 SFT -> RL

The worse it gets when we perform more SFT before RL

Thank You! Q&A

