

## A Experimental Details

This section describes the details of the experiment in Section 5. Our implementation, adapted from the Open-Reasoner-Zero codebase [3], is publicly available at [redacted].

### A.1 Hyperparameters

We use the same hyperparameters as those provided in Open-Reasoner-Zero [3], except that we reduce the number of samples per prompt from 64 to 32 to save computational resources. One training step proceeds as follows: we first randomly sample 128 prompts (rollout batch size) from the training dataset and generate 32 samples per prompt, totaling  $128 \times 32 = 4,096$  samples. We then divide the generated samples into 1 (12) training batch for the actor (critic), where each training batch is used for one optimizer update.

We tune the coefficient  $c$  in  $R_{\text{summary}}$  by searching over  $\{1e-4, 1e-3, 1e-2\}$ . Other Thinker-task-specific hyperparameters are selected using heuristics. We use a lower sampling temperature during summarization, as we observe that higher temperatures tend to produce less concise and consistent summaries.

For the baseline model, we use similar hyperparameters, except with a generation length of 8,000 tokens. We found that 8,000 tokens yield optimal baseline performance on R1.5B.

Table 2: Hyperparameters used in experiments.

Parameter	Value
<b>PPO</b>	
Rollout Batch Size	128
Number of Samples Per Prompt	32
Number of Epochs	1
Actor Learning Rate	1e-6
Number of Actor Update Steps	1
Critic Learning Rate	5e-6
Number of Critic Update Steps	12
Discount Rate $\gamma$	1
GAE Lambda $\lambda$	1
Clip Ratio $\epsilon$	0.2
KL Loss	None
Sampling Temperature	1
Sampling Temperature in Summarization	0.6
<b>Generation Length</b>	
Fast Thinking	1,000
Verification	6,000
Slow Thinking	6,000
Summarization	1,000
<b>Reward-specific</b>	
Coefficient $c$ in $R_{\text{summary}}$	1e-3
Minimum Length for Summarization	300

### A.2 Prompt Templates

The prompt templates used in the four stages of the Thinker task are illustrated in Box A.1. Note that not all prompts are necessarily used. For example, in training mode, if the agent’s fast answer is correct, the Slow Thinking and Summarization prompt will be skipped. Please refer to the main text for the termination conditions.

### Box A.1: Prompt Templates for Thinker Task

#### 1. Fast Thinking

**User:** Answer the below question with concise steps and output the final answer within `\boxed{}`. Limit your response below 1000 words.

This is the problem: {question}

**Assistant:** <Agent Response>

#### 2. Verification

**User:** Is your answer above correct? Please verify each step and the answer carefully. Output `\boxed{Yes}` if your answer is correct, or `\boxed{No}` if your answer is incorrect.

**Assistant:** <Agent Response>

#### 3. Slow Thinking

**User:** Your initial answer is incorrect. Now, think about the possible errors and consider alternative solutions. The reasoning process should be enclosed within `<think>...</think>`.

This is the problem: {question}

Let's think step by step and output the final answer within `\boxed{}`.

**Assistant:** <think> <Agent Response>

#### 4. Summarization

**User:** Your final answer is correct. Now summarize the steps leading to your final answer concisely and precisely, excluding internal reasoning. Limit your response between 300 and 1000 words.

This is the problem: {question}

**Assistant:** <Agent Response>

431

### 432 A.3 Computational Resources

433 Each training run for both the Thinker task and the baseline required approximately 7 days on two  
434 compute nodes, each equipped with 8 A100 GPUs. We use the Deepspeed [23], vLLM [24], and  
435 Ray [25] library for distributed training.

## B Result Details

This section describes additional experimental results that were omitted from the main text due to length constraints.

### B.1 Evaluation Results

Figure 7 and Figure 8 show the breakdown of Figure 4 from the main text, corresponding to the evaluation results of fine-tuning Q1.5B and R1.5B on the QA task or the Thinker task during training.

The detailed evaluation results of the ablated run in which the Summarization step is removed can be found in Table 3, labeled as SkipSum. Table 4 presents the standard error corresponding to the results reported in Table 1.

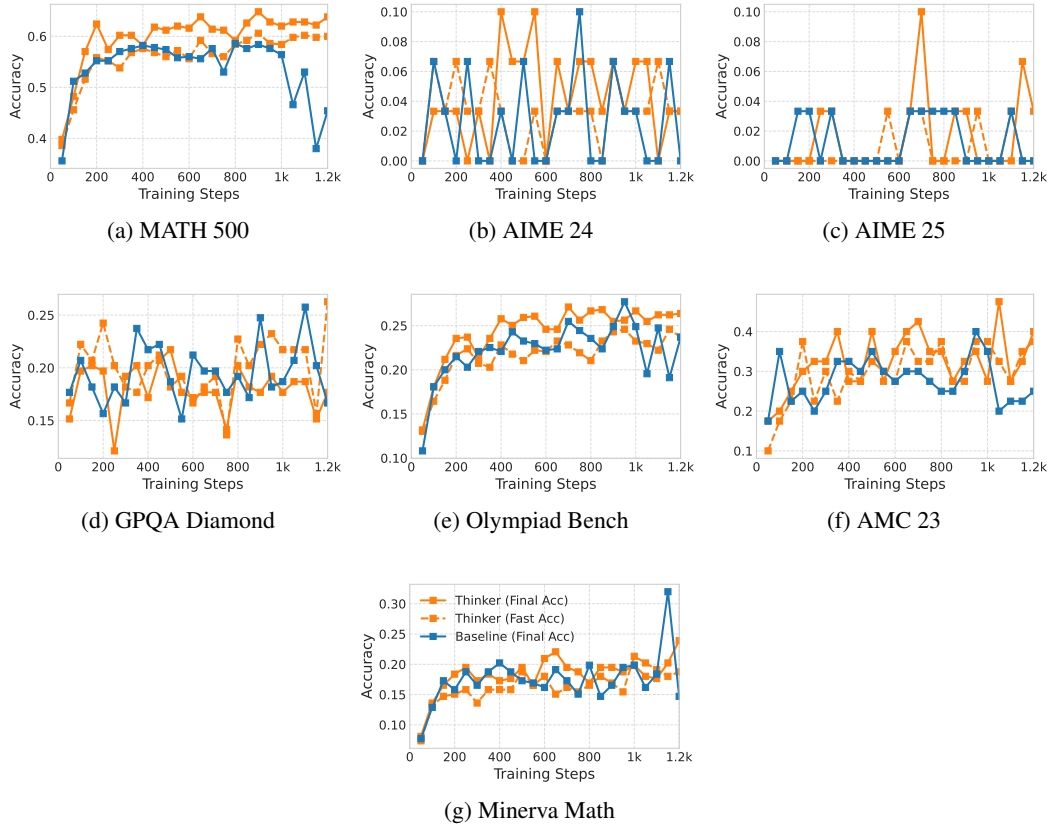


Figure 7: Detailed evaluation results of Q1.5B fine-tuned using QA task or Thinker task on individual mathematical reasoning benchmarks.

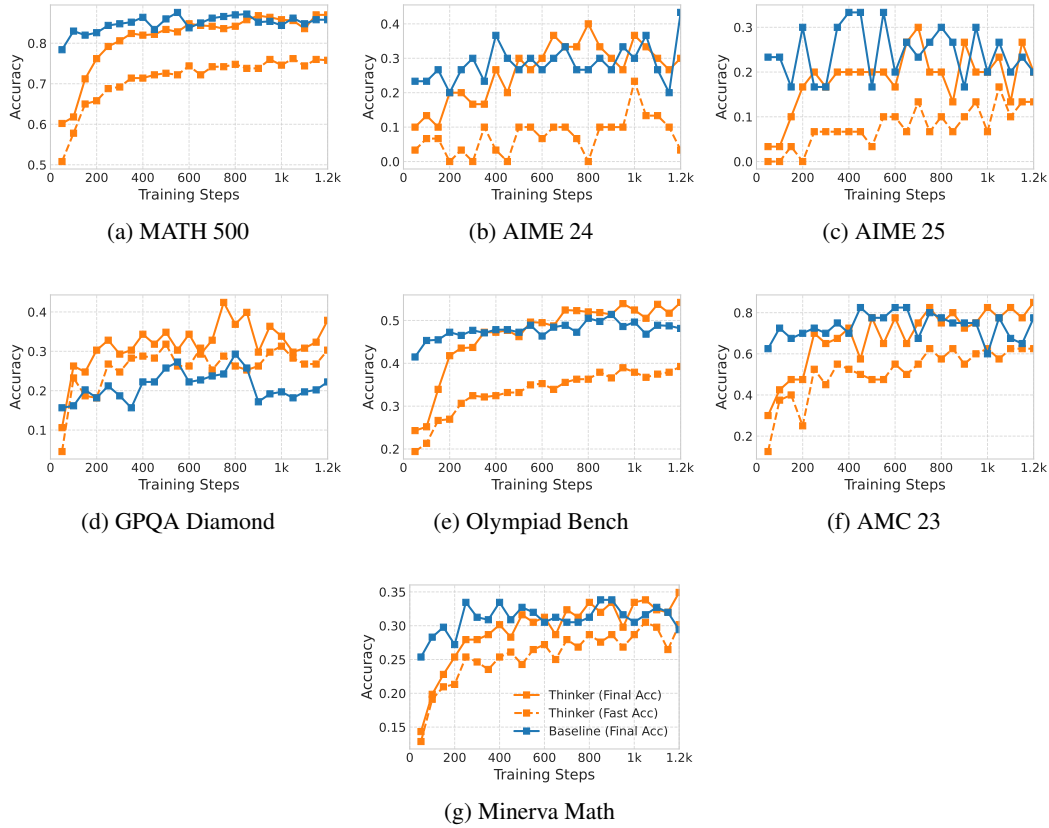


Figure 8: Detailed evaluation results of R1.5B fine-tuned using QA task or Thinker task on individual mathematical reasoning benchmarks.

Table 3: Mathematical reasoning performance of the Thinker agent trained without the Summarization stage. Average (Avg.) scores are presented. All scores are Pass@1 accuracy (%) averaged over 16 samples. Standard errors are provided in Table 4.

Method	MATH 500	AIME 2024	AIME 2025	GPQA Diamond	Olympiad bench	AMC 23	Minerva Math	College Math	Avg.
<b>Qwen2.5-1.5B (Q1.5B)</b>									
SkipSum	60.30	5.00	1.25	20.27	24.17	30.00	19.85	38.24	24.88
SkipSum-Fast	64.30	9.17	4.17	18.62	29.11	37.50	20.82	39.42	27.89

## B.2 Ablation Study on Fast Thinking Mode

To understand the importance of the Fast Thinking mode in the overall Thinker task, we experiment by using a less-trained agent to generate the Fast Thinking response, while still using the fully trained agent to generate responses for the remaining stages. This allows us to measure the impact of Fast Thinking quality on overall performance.

Specifically, we use four earlier R1.5B Thinker-agent checkpoints (Step 0, which is the pretrained model; Step 200; Step 400; and Step 600) to generate the Fast Thinking response, and use the fully trained R1.5B Thinker-agent for the remaining stages. We evaluate final accuracy across the eight benchmarks, as in the main evaluation. The results are shown in Figure 9.

We observe a general positive correlation between the Fast Thinking accuracy of a checkpoint and the final accuracy, suggesting that the Fast Thinking response has a substantial influence on subsequent stages. For instance, when we use the pretrained model (Step 0) to generate the Fast Thinking response, final accuracy drops significantly from 49.8% to 36.3%. However, we also observe that

Table 4: Standard error of performance on mathematical benchmarks in Table 1. All scores are in %.

Method	MATH 500	AIME 2024	AIME 2025	GPQA Diamond	Olympiad bench	AMC 23	Minerva Math	College Math	Avg.
<b>Qwen2.5-1.5B (Q1.5B)</b>									
Pretrained	0.36	0.00	0.00	0.40	0.15	0.91	0.19	0.09	0.14
Baseline	0.33	0.61	0.86	0.47	0.34	1.20	0.50	0.11	0.24
Thinker	0.19	0.67	0.57	0.39	0.09	1.32	0.13	0.08	0.17
Thinker-Fast	0.19	0.52	0.57	0.46	0.27	0.85	0.43	0.11	0.10
SkipSum	0.23	0.61	0.42	0.91	0.21	1.25	0.48	0.05	0.20
SkipSum-Fast	0.39	0.57	0.71	0.51	0.22	1.37	0.22	0.09	0.12
<b>DeepSeek-R1-Distill-Qwen-1.5B (R1.5B)</b>									
Pretrained	0.33	1.34	0.96	0.58	0.28	1.20	0.42	0.08	0.24
Baseline	0.24	1.05	0.52	0.56	0.34	0.99	0.30	0.08	0.22
Thinker	0.25	1.35	1.09	0.63	0.31	0.96	0.38	0.09	0.16
Thinker-Fast	0.25	1.10	0.91	0.45	0.21	1.46	0.24	0.07	0.24

458 this sensitivity diminishes as Fast Thinking performance improves. For example, using the Step  
 459 200 model, which has a moderate Fast Thinking accuracy of 28.9%, leads to a final performance of  
 460 49.18%—a minor drop from 49.8%.

461 We conjecture that this is due to the robustness of the Slow Thinking mode: since it is trained  
 462 specifically to handle incorrect Fast Thinking answers, it can often recover from slightly flawed  
 463 initial intuition. However, if the Fast Thinking intuition is very poor (as in the pretrained model), the  
 464 subsequent stages may fail to recover due to the lack of a meaningful starting point. A qualitative  
 465 analysis of how the Fast Thinking stage interacts with subsequent stages can be found in the case  
 466 study in Appendix C, which shows that the trained agent is able to correct flawed heuristics from the  
 467 Fast Thinking mode during the Verification and Slow Thinking mode.

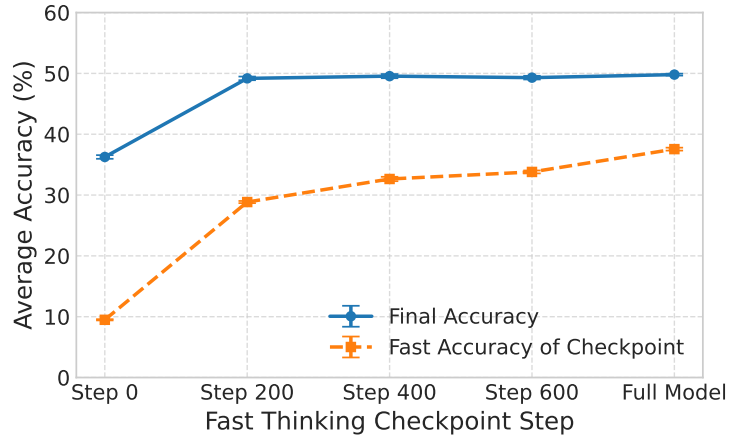


Figure 9: Final accuracy on the evaluation benchmarks of the Thinker agent (R1.5B) when its Fast Thinking stage is generated by model checkpoints at previous training steps. The original Fast Thinking accuracy of these respective checkpoints is also shown. All scores are Pass@1 accuracy (%) averaged over 16 samples. Error bars represent standard error, which are typically minor in this data.

## C Case Study

In this section, we present sampled responses from the fine-tuned R1.5B agent on the Thinker task, aiming to understand the behavior learned by the agent. Only responses with an incorrect fast answer are selected, so that the interaction between Fast Thinking and Slow Thinking can be observed.

### C.1 Case Study I: Identifying Flaws in Fast Thinking (Box C.1, Box C.2)

This example demonstrates the Thinker task’s ability to guide the agent from an uncertain, flawed initial attempt to a correct, verified solution by structuring its reasoning process.

**Fast Thinking.** The agent adopts a quick, System-1-like heuristic despite expressed uncertainty (“the side length of the larger hexagon is  $a + 2 \cdot 2 = a + 4$ ? Not sure.”) and proceeds with this flawed assumption ( $S = a + 4$ ). This leads to an incorrect perimeter,  $6\sqrt{3} - 12$ , which is physically implausible as it results in a negative value.

**Verification.** The agent directly confronts this flawed assumption. It explicitly questions the initial logic (“The side length of this larger hexagon might be  $s + 4$ ? Or is it  $s + 2$ ? Wait, reconsider. <...> the relationship isn’t straightforward.”). This critical re-evaluation leads to the correct geometric insight, yielding the relationship  $S = s + \frac{\sqrt{3}}{4}$ . Based on this corrected understanding, the agent identifies the error in its initial reasoning and conclusion, stating, “The previously given answer was  $6\sqrt{3} - 12$ . But that would not match. <...> Thus our initial approach is wrong.”

**Slow Thinking.** The agent then leverages the insight from Verification. It explicitly focuses on the “difference in the apothems” to re-derive  $S = s + \frac{\sqrt{3}}{4}$ . This demonstrates a clear adoption of the successful reasoning trace from Verification. The agent then systematically solves for the side length  $s$  and calculates the correct perimeter,  $18 - 4\sqrt{3}$ . Notably, it independently performs a numerical check, showcasing a deeper level of deliberation and confidence in its refined answer.

**Summarization.** The provided summary effectively distills the core mathematical steps for solving the problem into a clear, concise, and logically consistent sequence. It accurately establishes the relationship between the side lengths of the inner and outer hexagons, correctly formulates the equation for the path’s area, and finds the pool’s perimeter efficiently. Interestingly, it also employs a thinking block that reflects certain self-correction patterns observed in earlier steps.

This case highlights the agent’s capacity for targeted error identification and conceptual correction. The progression shows a clear refinement of reasoning, moving from the System-1-like heuristic in Fast Thinking to a more rigorous, System-2-like approach in Verification and Slow Thinking. The explicit references between stages—Verification critiquing Fast Thinking’s “initial approach” and Slow Thinking building directly on Verification’s apothem insight—underscore the efficacy of the structured task in fostering coherent, self-correcting thought processes.

### C.2 Case Study II: Propagation of Error from Verification to Slow Thinking (Box C.3)

This example provides a counterpoint to the previous successful error-correction cases. It demonstrates a scenario in which the agent arrives at an incorrect final answer due to error propagation and insufficient depth in later-stage reasoning.

**Fast Thinking.** The agent, faced with a complex product of fractional parts, makes a guess. After calculating the first few terms and noting the initial product starts with  $2 \cdot \frac{1}{2} = 1$ , it states: “Since initial term is 2 and product involves fractions potentially leading to  $1/2$ .” This leads to the incorrect answer  $D (\frac{1}{2})$ , driven more by the available options and a superficial pattern than a rigorous derivation.

**Verification.** This stage correctly identifies Fast Thinking’s answer as likely wrong. However, it introduces its own flawed heuristic: “Since each term  $\frac{a_k}{a_{k-1}}$  is less than 1, and multiple such terms multiplied, so perhaps the total product tends to zero.” While it correctly prompts a re-evaluation by concluding the original answer was incorrect, its reasoning toward an alternative answer (0) is based on a wrong heuristics without considering the specific structure of this sequence (which actually involves a telescoping product that simplifies to  $\frac{a_1}{a_{98}}$ , leading to a final answer of 1).

**Slow Thinking.** Instead of uncovering the correct algebraic simplification, Slow Thinking explicitly adopts and reinforces the flawed heuristic from Verification. After refining some calculations, it

517 reasons: “Alternatively, since the product seems to be approaching zero due to the product of many  
 518 terms each less than 1. But given the options, perhaps it’s 0.” This echoes Verification’s logic and  
 519 leads to the incorrect final answer B (0).

520 This case demonstrates how flawed reasoning can persist within the Thinker task. While Verification  
 521 correctly identified Fast Thinking’s initial guess as wrong, it unfortunately introduced its own plausible  
 522 but incorrect heuristic. Subsequently, Slow Thinking, anchored on this flawed premise inherited  
 523 from Verification, performing more calculations around it rather than challenging its validity. This  
 524 highlights the possibility of an anchoring bias on an incorrect heuristic, and ultimately, superficial  
 525 deliberation in the later stages that fails to correct the core flawed assumption.

### Box C.1: Example Response from Thinker Agent: Hexagon Path Problem

**Problem:** A path 2 m wide and with an area of  $36 \text{ m}^2$  has been built around a swimming pool in the shape of a regular hexagon. What is the perimeter of the pool?

**User:** <Fast Thinking Prompt: Answer concisely...>

**Assistant:**

<Initial Reasoning Trace showing uncertainty about side length>

3. <...> **the side length of the larger hexagon is  $a + 2 \cdot 2 = a + 4$ ? Not sure.**

<Calculations based on the assumption  $S = a + 4$ >

13. So  $8a = 8\sqrt{3} - 16 \Rightarrow a = \sqrt{3} - 2$ .

14. Perimeter of the hexagon is  $6a = 6(\sqrt{3} - 2) = 6\sqrt{3} - 12$ .

$$\boxed{6\sqrt{3} - 12}$$

**User:** <Verification Prompt: Is your answer above correct?>

**Assistant:**

<Reasoning Trace in Verification, questioning the  $S = s + 4$  assumption ( $a$  relabeled as  $s$ )>

**The side length of this larger hexagon might be  $s + 4$ ? Or is it  $s + 2$ ?**

Wait, reconsider. <...> the relationship isn’t straightforward.

<Derivation of the correct relationship  $S = s + \frac{4}{\sqrt{3}}$  based on apothems>

<Calculations using the corrected  $S$ , leading to Perimeter =  $18 - 4\sqrt{3}$ >

The previously given answer was  $6\sqrt{3} - 12$ .

But that would not match. So our approach seems incorrect.

<...> **Thus our initial approach is wrong.** Therefore, the original answer is incorrect.

<...>  $\boxed{No}$

**User:** <Slow Thinking Prompt: Your initial answer is incorrect. Now, think about errors...>

**Assistant:** <think>

<revisiting the core geometric error from Fast Thinking>

When you add a path around the hexagon, **the increase in side length can be found by considering the distance from the center.** <...>

The path is 2m wide, which I think refers to the distance from the pool to the path, that would mean the difference in the apothems is 2m. So:

$$\frac{S\sqrt{3}}{2} - \frac{s\sqrt{3}}{2} = 2$$

<Re-derivation of  $S = s + \frac{4}{\sqrt{3}}$ , consistent with Verification’s insight>

$$S = s + \frac{4}{\sqrt{3}}$$

<Systematic algebraic solution using the corrected  $S$ , leading to  $s = 3 - \frac{2\sqrt{3}}{3}$ >

<Calculation of perimeter =  $18 - 4\sqrt{3}$ , followed by numerical verification>

Hence, the perimeter is  $6s \approx 18 - 4\sqrt{3}$  meters. <...> Yes, that seems consistent.

Final Answer. The perimeter of the pool is  $\boxed{18 - 4\sqrt{3}}$  meters.

</think>

<Narrative solution in Slow Thinking, presenting the final corrected steps>

### Box C.2: Example Summary from Thinker Agent: Hexagon Path Problem

**Problem:** A path 2 m wide and with an area of  $36 \text{ m}^2$  has been built around a swimming pool in the shape of a regular hexagon. What is the perimeter of the pool?

**User:** <Summarization Prompt: Summarize concisely...>

**Assistant:**

<think>

1. Let  $s$  be the side length of the original hexagonal pool. The path surrounds the pool and forms a larger hexagon. The width of the path is 2m, so the side length of the larger hexagon is  $s + 2 \cdot (2/\sqrt{3}) = s + 4/\sqrt{3}$ . Wait, actually, the relationship is more complex.

2. The area of the path is the difference between the area of the larger hexagon and the smaller one. The area of a regular hexagon is  $(3\sqrt{3}/2) \cdot (\text{side length})^2$ .

3. Let  $s$  be the side length of the pool, then the side length of the larger hexagon is  $s + 2 \cdot (2/\sqrt{3}) = s + 4/\sqrt{3}$ . Wait, no, the correct relationship is that the side length increases by  $2 \cdot (2/\sqrt{3}) = 4/\sqrt{3}$ .

4. The area of the path is  $36 = (3\sqrt{3}/2) \cdot [(s + 4/\sqrt{3})^2 - s^2]$ .

5. Simplifying the equation, we find that  $s = 3 - (2\sqrt{3})/3$ . The perimeter is  $6 \cdot s = 18 - 4\sqrt{3}$ .

Thus, the perimeter of the pool is  $18 - 4\sqrt{3}$  meters.

</think>

Let  $s$  be the side length of the original hexagonal pool. The path width is 2m. The side length of the larger hexagon (pool + path) is  $S = s + \frac{4}{\sqrt{3}}$ . The area of the path is  $36 \text{ m}^2$ , which is the difference between the areas of the larger and smaller hexagons.

The area of a regular hexagon is  $\frac{3\sqrt{3}}{2} \times (\text{side length})^2$ . Therefore:

$$\frac{3\sqrt{3}}{2} \left( \left( s + \frac{4}{\sqrt{3}} \right)^2 - s^2 \right) = 36$$

Simplifying the term inside the brackets:

$$\left( s + \frac{4}{\sqrt{3}} \right)^2 - s^2 = s^2 + \frac{8s}{\sqrt{3}} + \frac{16}{3} - s^2 = \frac{8s}{\sqrt{3}} + \frac{16}{3}$$

Substituting this back into the area equation:

$$\frac{3\sqrt{3}}{2} \left( \frac{8s}{\sqrt{3}} + \frac{16}{3} \right) = 36$$

Multiplying through by  $\frac{3\sqrt{3}}{2}$ :

$$\begin{aligned} \frac{3\sqrt{3}}{2} \cdot \frac{8s}{\sqrt{3}} + \frac{3\sqrt{3}}{2} \cdot \frac{16}{3} &= 36 \\ 12s + 8\sqrt{3} &= 36 \end{aligned}$$

Solving for  $s$ :

$$\begin{aligned} 12s &= 36 - 8\sqrt{3} \\ s &= \frac{36 - 8\sqrt{3}}{12} = 3 - \frac{2\sqrt{3}}{3} \end{aligned}$$

The perimeter of the pool is  $6s$ :

$$6 \left( 3 - \frac{2\sqrt{3}}{3} \right) = 18 - 4\sqrt{3}$$

Thus, the perimeter of the pool is  $18 - 4\sqrt{3}$  meters.



### Box C.3: Example Response from Thinker Agent: Sequence Product Problem

**Problem:** Define the sequence  $\{a_n\} : a_1 = 1, a_2 = 2, a_{n+2} = a_n + a_{n+1}, n \in \mathbb{N}_+$ , then find  $\left[\frac{a_2}{a_1}\right] \cdot \left\{\frac{a_3}{a_2}\right\} \cdot \dots \cdot \left\{\frac{a_{99}}{a_{98}}\right\} \cdot \left[\frac{a_{98}}{a_2}\right]$ . Options: A. 2, B. 0, C. 1, D.  $\frac{1}{2}$

**User:** <Fast Thinking Prompt: Answer concisely...>

**Assistant:**

<Initial computation of sequence terms and first few product terms:  $[a_2/a_1] = 2, \{a_3/a_2\} = 1/2$ >

7. Notice that the product of fractional parts potentially telescopes or simplifies. However, it's complex.

8. Given the options A to D, and considering the pattern, likely answer is B.0 or C.1, but need to think carefully. **Since initial term is 2 and product involves fractions potentially leading to 1/2.**

D

**User:** <Verification Prompt: Is your answer above correct?>

**Assistant:**

<Re-evaluation of sequence and product structure, calculating  $[a_2/a_1] = 2$  and  $[a_{98}/a_2] \approx 17$ >

Wait, count the number of fractional parts: <...> 96 terms.

<...> **Since each term  $\{a_k/a_{k-1}\}$  is less than 1, and multiple such terms multiplied, so perhaps the total product tends to zero.**

Thus, the product is very small, so the overall expression would be close to zero. <...>

Therefore, the original answer was incorrect, so the correct answer should be No, as the initial answer was incorrect.

<...> **now I think the final answer might be B.0.**

**User:** <Slow Thinking Prompt: Your initial answer is incorrect. Now, think about errors...>

**Assistant:** <think>

<Recalculates sequence terms,  $[a_2/a_1] = 2$ . Identifies  $a_n = F_{n+1}$  and that  $F_{99}$  is even, refining  $[a_{98}/a_2]$  to  $F_{99}/2$ >

Given that each ratio  $a_k/a_{k-1}$  approaches the golden ratio  $\phi \approx 1.618$ .

Thus, each  $\{a_k/a_{k-1}\}$  approaches  $\phi - 1 = \psi \approx 0.618$ .

Thus, if each term  $\{a_k/a_{k-1}\}$  is about  $\psi \approx 0.618$ .

Thus, product of 96 terms [error in count, should be 97]:  $\psi^{96}$ .

Therefore, product =  $2 \cdot (\psi)^{96} \cdot \text{floor}(F_{99}/2)$ .

<...> **Alternatively, since the product seems to be approaching zero due to the product of many terms each less than 1. But given the options, perhaps it's 0.**

Therefore, the answer is B.0.

B

</think>

<Narrative solution, reinforcing the (incorrect) conclusion that the product is 0>

## D Preliminary Experiments on Larger Models

To understand the scalability of the Thinker task, we fine-tune DeepSeek-R1-Distill-Qwen-7B (R7B) [1], a larger version of R1.5B that also undergoes distillation using reasoning data from DeepSeek-R1. We use the same set of hyperparameters and training data as in the main experiments. The baseline corresponds to fine-tuning R7B on the standard QA task.

The average evaluation accuracy across the seven benchmarks (excluding CollegeMath due to the large number of questions) during training is shown in Figure 10. Due to limited computational resources, we have not yet trained the Thinker agent to convergence, but we already observe its performance surpassing that of the baseline, which has plateaued. The detailed performance of the best checkpoints from both runs can be found in Table 5.

Similar to the results observed in smaller models, we found that agents trained under the Thinker task consistently perform better than those trained on the baseline QA task across all benchmarks. The overall performance improved from 54.41% to 58.91%, representing a relative improvement of 8.27%. This suggests that larger models also benefit from the Thinker task. Additionally, we note that Thinker-Fast performance improves from 37.55% in R1.5B to 45.45% in R7B, demonstrating that the Fast Thinking mode scales well with model size.

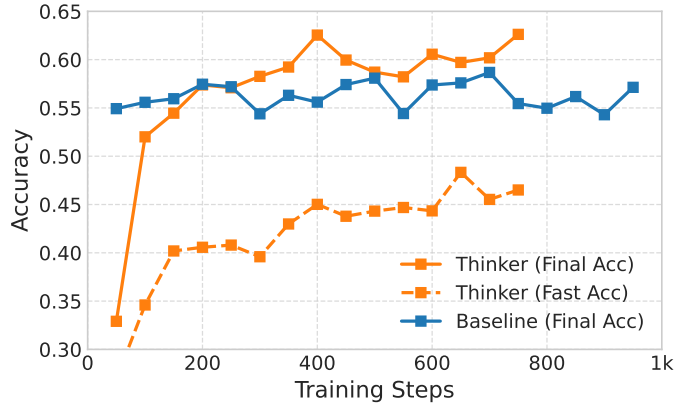


Figure 10: Evaluation performance of R7B averaged across seven common benchmarks.

Table 5: Performance comparison across various mathematical reasoning benchmarks. Average (Avg.) scores are presented. All scores are Pass@1 accuracy (%) averaged over 16 samples. Top score in each benchmark column is **bolded**. Standard errors are provided in Table 6.

Method	MATH 500	AIME 2024	AIME 2025	GPQA Diamond	Olympiad bench	AMC 23	Minerva Math	College Math	Avg.
<b>DeepSeek-R1-Distill-Qwen-7B (R7B)</b>									
Pretrained	84.05	37.50	28.54	17.58	37.92	36.41	34.49	40.72	39.65
Baseline	91.03	47.50	34.58	34.63	56.76	87.81	40.23	42.71	54.41
Thinker	<b>92.80</b>	<b>55.83</b>	<b>37.50</b>	<b>45.90</b>	<b>60.83</b>	<b>91.25</b>	<b>43.77</b>	<b>43.39</b>	<b>58.91</b>
Thinker-Fast	84.60	22.92	19.17	38.26	48.93	65.62	41.75	42.32	45.45

Table 6: Standard error of performance on mathematical benchmarks in Table 5. All scores are in %.

Method	MATH 500	AIME 2024	AIME 2025	GPQA Diamond	Olympiad bench	AMC 23	Minerva Math	College Math	Avg.
<b>DeepSeek-R1-Distill-Qwen-7B (R7B)</b>									
Pretrained	0.43	1.08	0.86	0.70	0.45	4.37	0.80	0.12	0.48
Baseline	0.19	1.62	1.29	1.00	0.31	0.79	0.27	0.07	0.39
Thinker	0.13	1.54	1.27	0.53	0.15	0.85	0.31	0.06	0.32
Thinker-Fast	0.25	0.52	0.57	0.67	0.19	1.24	0.36	0.06	0.21

## 545 **E Societal Impacts**

546 This research contributes to enhancing the reasoning capabilities of LLMs, which could positively  
547 impact areas like scientific advancement and education. By promoting more structured reasoning  
548 through the Thinker task, we aim for AI systems that are not only more performant but also potentially  
549 more understandable. However, as LLMs become more powerful, it remains essential to address  
550 potential risks, including misuse and unintended societal consequences, through continued research  
551 into AI safety, ethics, and governance.