
Learning to Reason Across Parallel Samples for LLM Reasoning

Jianing Qi¹, Xi Ye², Hao Tang^{1,3}, Zhigang Zhu^{1,4}, Eunsol Choi⁵

CUNY Grad Center¹, Princeton University², BMCC³, CCNY⁴,
New York University⁵

jqi@gradcenter.cuny.edu, xi.ye@princeton.edu, htang@bmcc.cuny.edu
zzhu@ccny.cuny.edu, eunsol@nyu.edu

Abstract

Scaling test-time compute brings substantial performance gains for large language models (LLMs). By sampling multiple answers and heuristically aggregate their answers (e.g., either through majority voting or using verifiers to rank the answers), one can achieve consistent performance gains in math domains. In this paper, we propose a new way to leverage such multiple sample set. We train a compact LLM, called Sample Set Aggregator (SSA), that takes a concatenated sequence of multiple samples and output the final answer, optimizing it for the answer accuracy with reinforcement learning. Experiments on five reasoning datasets demonstrate both the efficacy and efficiency of SSA. Notably, SSA improves over naive majority voting by 8% pass@5 on MATH. Furthermore, our 3B SSA surpasses model-based re-ranking with a much larger 72B process reward model. Our analysis also shows promising generalization ability of SSA, across sample set sizes, base model families and scales, and tasks. By separating LLMs to generate answers and LLMs to analyze and aggregate sampled answers, our approach can work with the outputs from premier black box models easily and efficiently.

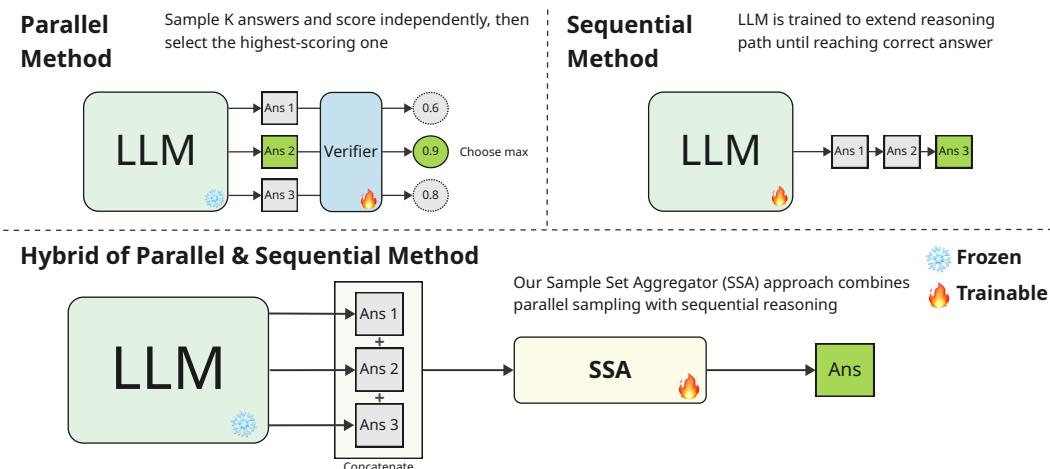


Figure 1: Illustration of our approach (bottom), parallel method (top left), and sequential method (top right). We train a compact LLM, called Sample Set Aggregator (SSA), to take a concatenated sequence of multiple samples and output the final answer.

1 Introduction

Recent advances in large language models (LLMs) have significantly enhanced their ability to perform complex reasoning tasks [El-Kishky et al., 2024, DeepSeek-AI et al., 2025]. Orthogonal to approaches to improve training LLMs through better learning objectives [Ouyang et al., 2022, Rafailov et al., 2023], architectures [Gu and Dao, 2024, Peng et al., 2023] or training dataset composition [Gunasekar et al., 2023, Wettig et al., 2025], recent work [Snell et al., 2025] explores a new dimension: test-time scaling. While having the LLM fixed, by allocating more computation at inference time (e.g., through repeated sampling and majority voting), one can improve the final task performance.

In general, test-time scaling methods fall into two paradigms. Parallel scaling generates multiple reasoning paths independently and aggregates them via strategies such as majority voting or best-of-N selection [Wang et al., 2023, Uesato et al., 2022, Lightman et al., 2024]. Sequential scaling, on the other hand, iteratively refines a single solution, often through prompting-based self-reflection or by incentivizing iterative computation [DeepSeek-AI et al., 2025, Muennighoff et al., 2025, Kumar et al., 2025]. In this paper, we introduce a novel test-time scaling approach that leverages both parallel and sequential scaling. Figure 1 illustrates our approach in comparison with prior work.

We separately have an LM to draft multiple answers and another LM to combine multiple answers to generate the final answer, naming the latter Sample Set Aggregator (SSA). SSA is optimized with reinforcement learning (RL) to maximize final answer accuracy. Unlike parallel scaling approaches that mostly view individual samples from LLMs in isolation, SSA can interpret multiple generations as representations of the LM’s output distribution, thus directly optimizing the synthesis of the final answer based on the landscape of the output distribution.

We conduct extensive experiments across five reasoning benchmarks under controlled test-time compute budgets. Results show that SSA substantially narrows the gap between actual model performance and oracle-best accuracy (pass@K), outperforming standard parallel strategies such as reward-based reranking. Moreover, we demonstrate that a compact SSA model can match the performance of reinforcement-trained larger models used in sequential scaling, suggesting the effectiveness of SSA as a lightweight way for sequential scaling. Further analysis highlights the generalization capabilities of SSA: SSA trained on one dataset for a particular model can successfully synthesize outputs from different model families and sizes across different tasks.

We summarize our key contributions and findings as follows:

- SSA, a lightweight LLM is introduced that concatenates K parallel candidates from a frozen base model and then performs one sequential RL step to generate the final answer. This single pass unifies the strengths of parallel and sequential test-time scaling, showing strong performance gain while training only a small model.
- Conceptually, we propose to *reason over the output distribution*. Specifically, instead of training the base LM, SSA optimizes over its sampled outputs. The base LM that produces answers remains to be a black box; SSA is trained only with sampled answers from base LMs.
- We observe broad and consistent empirical gains across five math benchmarks, two LLM families (Qwen 2.5, Llama 3.1) and three base sizes (7B, 14B, 32B), over strong baselines.

Project website: user074.github.io/ssa-parallel-reasoning

2 Related Work

Scaling test-time compute in parallel. Recent research has established the effectiveness of increasing compute used at inference time, known as test-time scaling [El-Kishky et al., 2024, DeepSeek-AI et al., 2025, Snell et al., 2025, Brown et al., 2025]. A prominent approach for test-time scaling focuses on parallel scaling, which samples multiple answers *independently* and aggregates them into a single answer. This aggregation can be performed through majority voting [Wang et al., 2023, 2024a] or more sophisticated selection mechanisms. For instance, some methods prompt language models to select from one of the multiple samples [Chen et al., 2024a], while others employ dedicated verifier models to score potential solutions [Cobbe et al., 2021, Uesato et al., 2022, Lightman et al., 2024, Li et al., 2023, Wang et al., 2024b] and take a weighted majority solution.

Beyond naively sampling multiple answers with fixed decoding strategy, researchers have explored advanced search strategies, such as beam search [Yao et al., 2023, Xie et al., 2023] and Monte-Carlo tree search (MCTS) [Li et al., 2025, Xie et al., 2024]. These search-based methods typically rely on verifiers [Xie et al., 2023] or process reward models [Cobbe et al., 2021, Wang et al., 2024b] to guide the decoding process, while still evaluating different rollouts independently. In contrast to these approaches that primarily assess samples in isolation, our approach learns to compare different samples jointly to determine the final answer.

Scaling test-time compute sequentially. Another line of research focuses on sequential scaling, which increases compute by iteratively updating and refining a solution. One way is to prompt or train LLMs to self-refine their proposed solutions [Madaan et al., 2023, Kumar et al., 2025, Qu et al., 2024, Chen et al., 2024b] in an iterative manner. Recent work has demonstrated the possibility of incentivizing LLMs to spend more tokens in a single completion through reinforcement learning [DeepSeek-AI et al., 2025, Team et al., 2025] or by forcing LLMs to continue their reasoning chains by appending "wait" token [Muennighoff et al., 2025]. While sequential scaling often yields performance improvements, recent studies debate on whether it outperforms simpler parallel scaling when controlling for compute [Zeng et al., 2025b, Hochlehnert et al., 2025a,b]. Our approach bridges these paradigms by operating on parallel samples but treating them as a sequence rather than as isolated instances. The growing length of CoT traces incurs substantial inference cost. Recent efforts have sought to improve efficiency by introducing length penalties [Aggarwal and Welleck, 2025, Sui et al., 2025], adaptive thinking [Fang et al., 2025, Zhang et al., 2025b, Lou et al., 2025], or applying early stopping [Zhang et al., 2025a, Yang et al., 2025b]. In contrast, our approach trains a compact aggregator that achieves both strong performance and efficiency.

Training language models for reasoning. More broadly, our work aims to enhance LLM performance on reasoning tasks, which has been a central pursuit in LLM development. Substantial efforts have been devoted to training LLMs as reasoning policies (distributions more likely to contain correct answers) through supervised fine-tuning on collections of chain-of-thought [Azerbaiyev et al., 2024, Puerto et al., 2024, Luo et al., 2023], or through reinforcement learning with rewards on intermediate steps [Uesato et al., 2022, Wang et al., 2024b, Kazemnejad et al., 2025] or based solely on final answer correctness [DeepSeek-AI et al., 2025]. Our research is more closely aligned with work on training LLMs as better verifiers for reasoning tasks and can complement solutions from any policy models. Existing methods train LLM verifiers to assess individual solutions using human annotations [Cobbe et al., 2021] or noisy labels derived from answer correctness [Wang et al., 2024b, Hosseini et al., 2024, Liang et al., 2024]. Our approach differs in that we train LLMs to verify sequences of solutions using reinforcement learning with a verifiable reward signal, the final answer correctness.

3 Method

3.1 Problem Formulation

Our method assumes two models: (1) LM_{ans} : a language model that generate a solution given an input, and (2) LM_{SSA} : a language model that takes an input and multiple solutions and generates the final solution.

Let \mathbf{x} be an input problem token sequence drawn from a data distribution \mathcal{D} , our method solves the problem with the following two steps:

Step 1: Parallel answer set generation. In this step, we use the LM_{ans} , which defines a conditional distribution $LM_{ans}(y|\mathbf{x})$ over all possible tokens $y \in Y$, to sample K candidate solution sequences from LM_{ans} , and obtain

$$Y_K = \{ \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_K \}, \quad y_i \sim LM_{ans}(\cdot | \mathbf{x}) \text{ independently.}$$

Step 2: Analyzing parallel answer set for the final answer. We use a separate trained language model LM_{SSA} that takes sampled answer sequence as input to generate the final answer \mathbf{y}_{final} .

$$\mathbf{y}_{final} \sim LM_{SSA}(\cdot | \mathbf{x}, Y_K)$$

Crucially, we treat the LM_{ans} as a black-box sampler of potential solutions, and train a much smaller model (LM_{SSA}) to do a post hoc aggregation as opposed to performing reinforcement learning on the answer model. Separating the answer generation model from the SSA introduces flexibility

in choosing a different generation model, and enable the use of larger models that are difficult to fine-tune due to limited compute or are only accessible via APIs. Thus we essentially propose a general conceptual framework for test-time computing.

To enable an LLM to function as SSA LM_{SSA} for synthesizing the final answer, we provide a natural language instruction directing the model to “think carefully and thoroughly evaluate the proposed answer, and identify one correct answer from the proposed candidates”. Following common practice in prior work [DeepSeek-AI et al., 2025], we specify a structured output format, which we use to extract the final answer. Please refer to Appendix B Figure 8 for details of the prompt.

In the rest of this section, we will introduce how we train the LM_{SSA} to maximize the correctness of its final answer. We consider two ways of optimizing the SSA: reinforcement learning (§ 3.2) and supervised-finetuning (§ 3.3).

3.2 Training SSA with Reinforcement Learning

Let \mathbf{y}^* be the gold solution for \mathbf{x} . For the final \mathbf{y}_{final} given by SSA $LM_{SSA}(\cdot | \mathbf{x}, Y_K)$, we design a verifiable reward $R(\mathbf{y}_{final}, \mathbf{y}^*)$ that mainly verifies the answer correctness, and update LM_{SSA} to maximize the expected reward:

$$\mathbb{E}_{[(x, y^*) \sim \mathcal{D}, Y_K \sim LM_{ans}(\cdot | x)]} [R(LM_{SSA}(x, Y_K), \mathbf{y}^*)]. \quad R(\mathbf{y}_{final}, \mathbf{y}^*) = \begin{cases} 1, & \text{if } \mathbf{y}_i \text{ is correct,} \\ 0.05, & \text{if only format is correct,} \\ 0, & \text{otherwise.} \end{cases}$$

Following prior work [DeepSeek-AI et al., 2025], our reward also consider the format of the output specified in the prompt. Specifically, we let the reward R to be 1.0 if the answer is correct; we let the reward be 0.05 if the output follows the format and is incorrect; we let the reward to be 0.0 if the output does not follow and format.

We use the Group-Relative Policy Optimization (GRPO [Shao et al., 2024]) as our optimization algorithm, which simplified value function of PPO [Schulman et al., 2017] with a normalized reward from a group. For the convention of notation, we use π_θ for LM_{SSA} with parameter θ . For completeness, we describe it here. It maximizes:

$$J_{GRPO}(\theta) = \mathbb{E}_{\substack{Y_K \sim LM_{ans}(\cdot | \mathbf{x}) \\ \{\mathbf{y}_i\} \sim \pi_{\theta_{old}}(\cdot | \mathbf{x}, Y_K)}} \frac{1}{G} \sum_{i=1}^G \frac{1}{|\mathbf{y}_i|} \sum_{t=1}^{|\mathbf{y}_i|} \left[\min(\rho_{i,t}(\theta) \hat{A}_{i,t}, \text{clip}(\rho_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon) \hat{A}_{i,t}) \right] - \beta D_{KL}(\pi_\theta \| \pi_{ref}),$$

with importance weight $\rho_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} | \mathbf{x}, Y_K, y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t} | \mathbf{x}, Y_K, y_{i,<t})}$ with the advantage as $\hat{A}_{i,t} = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$, Reward as $r_i = R(\mathbf{y}_i, \mathbf{y}^*)$, $\mathbf{r} = r_1, r_2, \dots, r_n$, and ε, β are hyperparameters. This policy gradient method is to maximize the $\hat{A}_{i,t}$ which translates to expected reward from $r_i = R(\mathbf{y}_i, \mathbf{y}^*)$.

Implementation We use a training set combining GSM8K and MATH [Cobbe et al., 2021, Hendrycks et al., 2021]. To prepare the dataset, we use Qwen2.5-7B-Instruct model as LM_{ans} to generate five answers for each questions [Qwen et al., 2025]. Then we concatenate the question and each answer (ordered randomly) as input context of SSA LLM. We filter out instances where the combined answers exceed 4000 tokens, or where fewer than three valid answers are available. This results a dataset with a total size of 17.4k. We set the GRPO sample size to be 8, resulting in $8 * 17.4k \approx 140k$ during training.

3.3 Training SSA With Supervised Finetuning

We also explore the training of the SSA via supervised fine-tuning (SFT), leveraging a stronger model to construct oracle reasoning paths across multiple sampled answers to identify correct final answer. Once we constructed the oracle dataset, we use the standard language modeling objective, but only training on the output part. We assume a dataset of oracle reasoning path that leads to \mathbf{y}^* . If $\mathbf{y}^* = (y_1^*, y_2^*, \dots, y_T^*)$ is tokenized into T tokens, the SFT loss is to minimize:

$$\mathcal{L}_{SFT}(\theta) = - \sum_{(\mathbf{x}, Y_K, \mathbf{y}^*) \in \mathcal{D}} \sum_{t=1}^T \log [LM_{SSA}(y_t^* | \mathbf{x}, Y_K, \mathbf{y}_{<t}^*)],$$

Implementation We prompt Qwen 2.5 7B Instruct model for 5 candidate solutions per question in the GSM8K dataset. Then, we provide concatenated candidate solutions to GPT-4.1 Nano model along with the original question and ground-truth answer (The exact prompt is in Appendix B). GPT-4.1 Nano then provides a step-by-step reasoning process to identify and generate the best final answer from these candidates, yielding 7.47k training examples. Overall, this approach achieves a 96.24% match rate with the original ground-truth answers.¹

4 Experiments

4.1 Experimental Settings

Datasets. For evaluation, we use an array of commonly used math reasoning datasets: the test split of GSM8K, MATH as the in-domain evaluation sets, and AIME 2024 [MAA, 2024], AMC 2023, and Olympiad [He et al., 2024] as the test sets. We use the extracted answers and grade them against the ground truth answers, using the library that has been used in prior work [Lightman et al., 2024].

Base Models for Candidate Generation (LM_{ans}). We use the Qwen-2.5-Instruct model, with sizes 7B, 14B, and 32B to generate K of answers. We use a decoding temperature 0.5 to construct the training and test dataset. In the training phase, we use $k = 5$ to train the SSA model. In the testing phase, we evaluated the performance with $k = \{5, 10, 15\}$.

Base Models for SSA (LM_{SSA}). We use the Qwen-2.5-base model with sizes 0.5B, 1.5B, and 3B. We chose the Qwen-2.5 model family due to its popularity and wide availability for PRM verifiers, making it possible to compare against existing PRM verifiers from the same model family.

Training Implementation Details. For the training library, we use torchtune due to its efficient VRAM management [torchtune maintainers and contributors, 2024]. We use GRPO batch sample group size 8, batch size 1, temperature 1.0, AdamW optimizer, KL coefficient 0.01, and learning rate 1e-5. We trained all experiments one epoch. For hardware, we use 8*H100 80GB for training. For shorter context, it is also possible to train with 48GB VRAM.

4.2 Comparison Systems

The following models are compared, including the proposed SSA with three variations.

Rule-based Baselines. We report pass@ k , which reports the percentage of examples where any of k number of solution is correct [Brown et al., 2025]. Pass@ k also serves as the oracle performance assuming we have an oracle verifier. We also report the **majority vote** [Wang et al., 2023], counting the most frequent answer among the answer set.

Outcome Reward Model (ORM). We use off the shelf Llama-3.1-8B ORM model trained with RLHF-Reward Modeling [Xiong et al., 2024] to re-rank multiple samples, selecting the best one scored by the reward. We will call it Llama-ORM (8B). It is trained with 273k data. For ORM evaluations, it adds an evaluation token at the end of the answer and the model will provide a score.

Process Reward Model (PRM). One prominent way to leverage multiple parallel sample is using process reward models. We use the Qwen-7B PRM (Qwen PRM) and Qwen-72B PRM model from Zhang et al. [2025c] to re-rank the candidate solutions.

We note that compared to the PRM model, our SSA is trained with significantly less resources regarding both model scale (0.5B - 3B vs 7B) and the amount of training data. Specifically, Qwen-7B PRM is trained with more than 500,000 queries with 6 to 8 answers with step labels, resulting over 3 million total training data (about 20 \times larger than the training data size of SSA). For a more fair comparison, we also include Shepherd PRM [Wang et al., 2024b], 7B model trained with 440k total step level data from their MCTS roll out. Additionally, Qwen PRM initialized from Qwen2.5-Math-7B-Instruct [Yang et al., 2024], whereas we initialized from the Qwen 2.5 base model.

For evaluation, we follow the training method of Qwen PRM [Zhang et al., 2025c], and we separate steps with ‘\n\n’ delimiter. Then we use PRM model to compute a score for each step and use their product to calculate the response score since this yielded the best performance in their experiments.

¹When we inspected the remaining cases, we find many ground-truth labels are incorrect.

Table 1: Results (accuracy %) with $k = 5$ candidate answers generated by Qwen2.5-7B-Instruct as LLM_{ans} model. Aggregation overheads (seconds/question) is measured end to end with AMC23 40 questions with $k=5$.

Method	Aggregation Overhead (s)	Datasets					Avg
		GSM8K	MATH	AIME24	AMC23	Olympiad	
pass@1	-	89.01	64.00	10.00	37.50	27.00	45.50
pass@5	-	95.45	78.00	16.67	67.50	41.10	59.74
Majority Vote	-	91.66	68.20	10.00	47.50	31.01	49.67
USC w/ Qwen 3B	8.48	61.18	42.80	6.67	17.50	15.43	28.72
USC w/ Qwen 7B	5.89	5.89	61.20	6.67	47.50	28.78	43.83
LLama ORM (8B)	0.61	93.1	67.40	13.33	47.50	28.64	49.99
Qwen PRM (7B)	0.64	92.57	69.40	13.33	57.50	32.05	52.97
Qwen PRM (72B)	5.86	92.87	69.6	13.33	57.50	33.68	53.40
Shepherd PRM (7B)	0.65	90.75	64.40	13.33	35.00	27.00	46.10
SSA RL (0.5B)	0.20	92.65	75.40	10.00	57.50	37.98	54.71
SSA RL (1.5B)	0.33	92.49	76.60	10.00	52.50	38.72	54.06
SSA RL (3B)	0.55	93.25	76.80	13.33	57.50	39.76	56.13

Universal Self Consistency (USC) [Chen et al., 2024a]. USC is a prompting-based method takes a concatenation of the multiple parallel sample answers, and instructs a LM to generate the final answer. The exact prompt can be found in Appendix B. The task setting is equivalent to our SSA method, but the LM is not trained to optimize for the final answer.

SSA and Variations. We train three types of the SSA models, one trained with the SFT objective only (SFT), one trained with SFT objective and then with the RL objective (SFT + RL), and one that is trained with RL objective only (RL). For SFT objective, we use the dataset described in Section 3.3 to finetune the model. We use learning rates 5e-5, 2e-5, and 1e-5 for 0.5B, 1.5B, and 3B correspondingly. For RL objective, we use GRPO method described in the Section 3.2. For SFT + RL objective, we use the trained model from the SFT objective, then we continue to train the RL objective on top of it. For each variant, we present models of three different sizes, trained from Qwen-0.5B, 1.5B, and 3B base models checkpoint respectively.

As an ablation for the SSA (RL), we also present SSA trained on GSM8K data only. Another ablation is to train the SSA RL version without thinking. It helps us to understand how each design decisions affect the output and performance under controlled settings.

5 Main Results

5.1 Comparison with Parallel Scaling Baselines

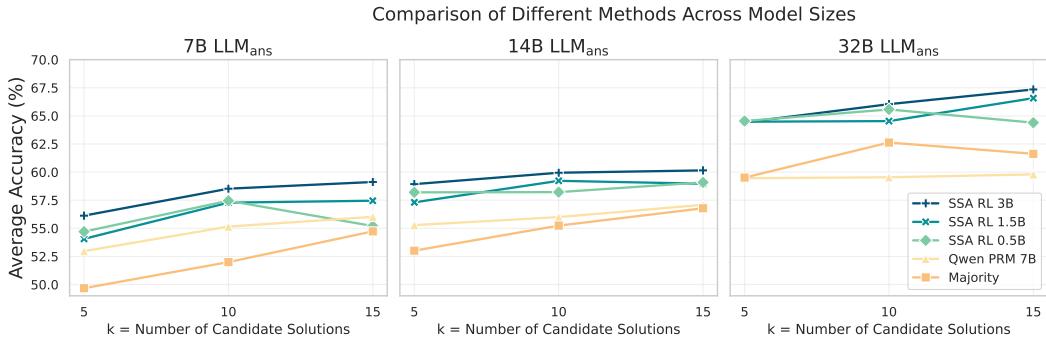


Figure 2: Compare the performance of SSA RL, PRM, and Majority Vote methods across Qwen 2.5 LLM_{ans} model sizes (7B, 14B, 32B) and number of candidate solutions $k = 5, 10, 15$.

We report performances across GSM8K, MATH, AIME24, AMC23, Olympiad benchmark in Table 1, more detailed results can be seen in Table 4 Appendix A.

Table 2: Generalization results for different model families (accuracy %) with $k = 5$ candidate from Llama 3.1 8B Instruct LLM_{ans} models. Aggregation overhead (seconds) is measured end to end with AMC23 40 questions with $k=5$.

Method	Aggregation Overhead (s)	Datasets					Avg
		GSM8K	MATH	AIME24	AMC23	Olympiad	
Majority Vote	-	87.95	50.40	6.67	35.00	17.66	39.54
USC w/ Qwen 7B	5.89	84.15	51.6	6.67	37.5	19.29	39.84
LLama ORM (8B)	0.61	89.61	51.8	13.33	30.00	18.4	40.63
Qwen PRM (7B)	0.64	91.51	56.20	13.33	35.00	20.77	43.36
SSA RL (0.5B)	0.20	88.17	52.80	10.00	30.00	20.18	40.23
SSA RL (1.5B)	0.33	88.48	56.60	10.00	27.50	20.47	40.61
SSA RL (3B)	0.55	89.08	57.80	10.00	32.50	20.62	42.00

Overall, the results suggest that SSA is highly effective at leveraging multiple sampled solutions. In comparison to baseline methods, SSA substantially narrows the performance gap relative to the oracle (pass@5). Notably, even the smallest SSA variant leads to strong gains. SSA (0.5B) achieves an average performance of 54.7%, even outperforming the Qwen PRM (7B) baseline which uses a much larger model. We focus on SSA RL results in this section and leave more detailed results including SSA SFT and SSA SFT + RL in Appendix A.

5.2 Generalization Capabilities of SSA.

Generalization across scales. SSA trained on outputs from a smaller answer model (Qwen 2.5 7B Instruct) can generalize to outputs from larger answer models (14B, 32B). Figure 2 compare the overall performance of SSA and the other baselines under different numbers of samples. We can see that SSA RL consistently outperform the baselines under different number of candidate solution samples across Qwen models with scales ranging from 7B to 32B. More SSA versions and results are in Table 4 in the Appendix A

Generalization across model families. In addition to the Qwen 2.5 7B Instruct model, we also tested Llama 3.1 8B Instruct model for inference [Grattafiori et al., 2024].

Results are in Table 2. We can see the SSA method can also generalize well outside of the training data distribution and outperforming Majority Vote, USC, and ORM methods. While Qwen PRM performs slightly better than SSA RL 3B here (+1.36%), it requires substantially more training data (over 3 million examples, x21 times) and a larger model (7B).

Generalization to harder datasets. In Table 1, while we have trained the SSA models with problems only from GSM8K and MATH, we observe substantial performance gains on other harder datasets (AMC23 and Olympiad). Such results indicate SSA can **generalize to unseen test sets** outside of the training datasets.

In addition, we tested SSA on general tasks (specifically ARC-C, MMLU-Pro, and TruthfulQA [Clark et al., 2018, Wang et al., 2024c, Lin et al., 2022]) without additional training. We observe minor gains compared to majority vote, especially when using both SFT and RL, in ARC-C and MMLU-Pro dataset, but very little for TruthfulQA dataset. Further research is needed to study generalization across different domains. More details can be found in Appendix A.8.

5.3 Comparison with Sequential Test-time Scaling

We have demonstrated the advantageous of SSA over diverse parallel scaling methods in early §5.1. We now compare the performance on SSA and reasoning models that are enabled to spend more test-time compute sequentially with an RL-based approach as in [DeepSeek-AI et al., 2025]. In particular, we use the results reported by Simple-RL Zoo from Zeng et al. [2025a] because they also uses Qwen-series models and train them on GSM8K and MATH dataset. In our comparison, we match the test-time computation, specifically, our 3B SSA is trained with Qwen-2.5-7B answers, but we apply it on top of 7B, 14B, 32B LLM_{ans} models. We sample 8 candidate solutions from the LLM_{ans} model with a maximum token length of 1,024 per answer, roughly matching the maximal generation length, 8,192, of these reasoning models.

Table 3: Performance of SSA (accuracy, % averaged over 5 datasets) under increasing number of samples K . * denote runs that use the adapted two-stage SSA (applied when $K \geq 32$). SSA at $K = 15, 64$ even outperform majority voting and PRM at $K = 128$.

Answer K	15	32	64	128
Majority Vote	54.73	54.07	54.43	55.11
Qwen PRM (7B)	56.02	55.37	58.49	56.58
SSA RL (3B)	59.12	58.42*	59.78*	58.68*

As shown in Figure 3, both the SSA and the RL training from sequential scaling improve over majority voting. While SSA slightly underperforms sequential scaling with RL training, note that we only optimize the 3B model as opposed to the 7B, 14B, and 32B model.

5.4 Scaling SSA to handle a larger sample set

In this section, we study SSA performance at larger K values ($K=32-128$). So far, SSA considers all K answers as a concatenated single input. However, concatenation at larger K quickly runs into context length limitations, leading to performance degradation or failures in answer extraction. For instance, with $K = 32$, the concatenated input length approaches $\sim 30K$ tokens, nearly saturating the 32K context window of our backbone model (Qwen-2.5-3B). Furthermore, prior work highlights that the effective usable window is often shorter than the nominal size [Yang et al., 2025a, Liu et al., 2024, Hsieh et al., 2024, Ye et al., 2025].

Two-Stage SSA To address this, we introduce a simple yet effective two-stage adaptation of SSA for large K similar to [Toshniwal et al., 2025]. In the first stage, we evenly split the K samples into l_2 groups of size l_1 , and run SSA independently on each group to produce l_2 intermediate candidates. In the second stage, SSA is applied to aggregate these l_2 candidates into a final answer. Instead of running SSA once, with this two-stage adaptation, we run SSA for $l_2 + 1$ times, l_2 times to aggregate l_1 inputs, and once to aggregate l_2 inputs. We include full descriptions of two stage SSA in Appendix A.5. For our experiments, we set the hyperparameter $l_1 = 15$. This sets $l_2 = 3$ for $K = 32$, and $l_2 = 5$ for $K = 64$ and $l_2 = 9$ for $K = 128$.

Results Table 3 presents results of SSA under increasing K on Qwen-2.5-7B outputs, compared against majority voting and PRM-7B. We find that SSA makes more efficient use of samples, achieving strong performance. For example, SSA at $K = 15$ already surpasses both majority vote and PRM at $K = 128$. By contrast, PRM shows less stable performance (with noticeable fluctuations across K), as it operates by scoring individual answers without cross-sample aggregation, making it more vulnerable to the noisy reward model.

6 Analyses

6.1 Comparing RL vs SFT

As an alternative to RL, we can train the SSA via supervised fine-tuning (SFT). How would SFT compare to RL? In our setup, we use the dataset described in Section 3.3 and fine-tune 0.5B, 1.5B, and 3B SSA models on this data for one epoch (learning rates 5e-5, 2e-5, and 1e-5, respectively). For comparison, we also train SSA models with *only* RL (on the same GSM8K problems for SFT) for controlled settings. Figure 4a and Table 6 (Appendix A.2) summarize the results.

We observe that SFT slightly outperforms RL in 0.5B, suggesting that direct supervision from high-quality data can help small model. For larger models (1.5B, 3B), RL yields better accuracy and robustness to larger k during inference. SFT trained on LLM_{ans} with $k = 5$ generalizes less

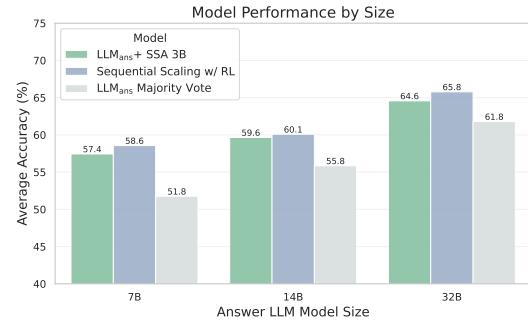
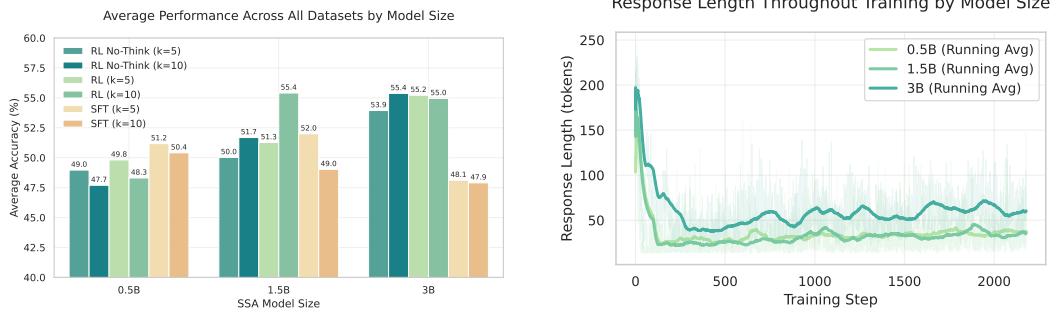


Figure 3: Performance comparison between sequential scaling with RL and our SSA



(a) Performance comparison of different training methods (SFT, No-Think, RL) across model sizes.

(b) Response length evolution during training for different model sizes.

Figure 4: Training method performance and response length analysis. (a) Average accuracy across datasets shows RL method is more generalizable than SFT method, with performance improving for larger models. (b) Response length trends during training show a rapid decrease of output length.

effectively to $k = 10$. Overall, both SFT and RL can work well, but SFT’s performance relies heavily on dataset quality and alignment. We also note that SFT tends to produce more readable reasoning traces, whereas RL outputs are often minimal (example outputs can be found in Appendix C).

6.2 To think or not to think

A distinct pattern we noted in our RL training is the reduction of the thinking tokens. Figure 4b shows how response length quickly drops during training across all model sizes. The model often simply repeats the provided instruction format for thinking (e.g., ‘<think>reasoning process here</think>’) followed immediately with the final answer. Examples are shown in Appendix C Figure 14. This contrasts with other RL-based reasoning models that generate longer explanations with more training [DeepSeek-AI et al., 2025, Zeng et al., 2025a]. It is likely because our SSA is conditioned on multiple candidate solutions, reducing the utility of detailed reasoning.

To assess the necessity of explicit reasoning tokens, we train an RL model variant without the reasoning step (‘No Think’), where the model directly generates the answer (see Appendix B for prompt and Appendix C Figure 15 for example). Figure 4a summarizes the results, and Table 6 in Appendix A.2 compares this variant with the original approach across three model sizes in details. We observe minor performance degradation without explicit thinking, suggesting detailed reasoning tokens might not substantially contribute to the final performance in our current setup.

7 Conclusion

We introduce SSA, a small LM trained with RL that can leverage outputs from a larger base LLM. By decoupling RL training from base model, it suggests that the quality of the base model knowledge is more important for performance. This novel hybrid approach, blending parallel and sequential scaling methods, provides practical benefits for plug and play.

Limitations Across benchmarks, SSA succeeds mainly by picking an correct candidate among input candidate answers. Failures happens mainly when the gold answer is absent among the candidates (Appendix A.6 quantifies this trend). We experimented with enabling SSA to generate new final answers by cutting the last 10% of candidate answer tokens, but this does not yield better performance (See Appendix A.7 for details).

Future work We outline few possibilities for further improving and extending our approach. Promising directions include scaling the number of outputs to be aggregated, as well as building SSA that can incorporate outputs from multiple LLMs. Improving its performance for diverse application beyond mathematical reasoning, as well as enhancing SSA’s new target answer synthesis ability can be fruitful. Overall, we believe SSA’s hybrid approach offers a promising direction for future research in LLM reasoning.

Acknowledgments

The work is supported by the National Science Foundation (NSF) through Awards #2131186 (CISE-MSI), #1827505 (PFI), and the US Air Force Office of Scientific Research (AFOSR) via Award #FA9550-21-1-0082. The work is also supported by a College-wide Research Vision (CRV) Fund from the CCNY Provost’s Office and the Google CyberNYC Initiative. This work used Google Cloud through the CloudBank project, which is supported by NSF Award #1925001. We gratefully acknowledge use of the research computing resources of the Empire AI Consortium, Inc, with support from Empire State Development of the State of New York, the Simons Foundation, and the Secunda Family Foundation [Bloom et al., 2025].

References

P. Aggarwal and S. Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning, 2025. URL <https://arxiv.org/abs/2503.04697>.

Z. Azerbayev, H. Schoelkopf, K. Paster, M. D. Santos, S. M. McAleer, A. Q. Jiang, J. Deng, S. Biderman, and S. Welleck. Llemma: An open language model for mathematics. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=4WnqRR915j>.

S. Bloom, J. C. Brumberg, I. Fisk, R. J. Harrison, R. Hull, M. Ramasubramanian, K. V. Vliet, and J. Wing. Empire AI: A new model for provisioning AI and HPC for academic research in the public good. In *Practice and Experience in Advanced Research Computing (PEARC '25)*, page 4, Columbus, OH, USA, July 2025. ACM. doi: 10.1145/3708035.3736070. URL <https://doi.org/10.1145/3708035.3736070>.

B. Brown, J. Juravsky, R. S. Ehrlich, R. Clark, Q. V. Le, C. Re, and A. Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2025. URL <https://openreview.net/forum?id=0xUEBQV54B>.

X. Chen, R. Aksitov, U. Alon, J. Ren, K. Xiao, P. Yin, S. Prakash, C. Sutton, X. Wang, and D. Zhou. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*, 2024a. URL <https://openreview.net/forum?id=LjsjHF7nAN>.

X. Chen, M. Lin, N. Schärli, and D. Zhou. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations*, 2024b. URL <https://openreview.net/forum?id=KuPixIqPiq>.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You,

Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.

A. El-Kishky, D. Selsam, F. Song, G. Parascandolo, H. Ren, H. Lightman, H. W. Chung, I. Akkaya, I. Sutskever, J. Wei, J. Gordon, K. Cobbe, K. Yu, L. Kondraciuk, M. Schwarzer, M. Rohaninejad, N. Brown, S. Zhao, T. Bansal, V. Kosaraju, and W. Zhou. Learning to reason with llms, Sep 2024. URL <https://openai.com/index/learning-to-reason-with-llms/>.

G. Fang, X. Ma, and X. Wang. Thinkless: Llm learns when to think. *ArXiv*, abs/2505.13379, 2025.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Srivankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Couder, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenber, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, G. Guangyi, G. Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani,

M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=tEYskw1VY2>.

S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. D. Giorno, S. Gopi, M. Javaheripi, P. Kauffman, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, H. S. Behl, X. Wang, S. Bubeck, R. Eldan, A. T. Kalai, Y. T. Lee, and Y. Li. Textbooks are all you need, 2023. URL <https://arxiv.org/abs/2306.11644>.

C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

A. Hochlehnert, H. Bhatnagar, V. Udundarao, S. Albanie, A. Prabhu, and M. Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility, 2025a. URL <https://arxiv.org/abs/2504.07086>.

A. Hochlehnert, H. Bhatnagar, V. Udundarao, S. Albanie, A. Prabhu, and M. Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility, 2025b. URL <https://arxiv.org/abs/2504.07086>.

A. Hosseini, X. Yuan, N. Malkin, A. Courville, A. Sordoni, and R. Agarwal. V-STAR: Training verifiers for self-taught reasoners. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=stmqBSW2dV>.

C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, and B. Ginsburg. Ruler: What's the real context size of your long-context language models?, 2024. URL <https://arxiv.org/abs/2404.06654>.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020. URL <https://arxiv.org/abs/2001.08361>.

A. Kazemnejad, M. Aghajohari, E. Portelance, A. Sordoni, S. Reddy, A. Courville, and N. L. Roux. VinePPO: Unlocking RL potential for LLM reasoning through refined credit assignment, 2025. URL <https://openreview.net/forum?id=5mJrGtXVwz>.

A. Kumar, V. Zhuang, R. Agarwal, Y. Su, J. D. Co-Reyes, A. Singh, K. Baumli, S. Iqbal, C. Bishop, R. Roelofs, L. M. Zhang, K. McKinney, D. Shrivastava, C. Paduraru, G. Tucker, D. Precup, F. Behbahani, and A. Faust. Training language models to self-correct via reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=CjwERcAU7w>.

S. Li, S. Dong, K. Luan, X. Di, and C. Ding. Enhancing reasoning through process supervision with monte carlo tree search. *arXiv preprint arXiv:2501.01478*, 2025.

Y. Li, Z. Lin, S. Zhang, Q. Fu, B. Chen, J.-G. Lou, and W. Chen. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023.

Z. Liang, Y. Liu, T. Niu, X. Zhang, Y. Zhou, and S. Yavuz. Improving LLM reasoning through scaling inference computation with collaborative verification, 2024. URL <https://openreview.net/forum?id=Qyile3DctL>.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=v8L0pN6EOi>.

S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL <https://arxiv.org/abs/2109.07958>.

N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL <https://aclanthology.org/2024.tacl-1.9/>.

C. Lou, Z. Sun, X. Liang, M. Qu, W. Shen, W. Wang, Y. Li, Q. Yang, and S. Wu. Adacot: Pareto-optimal adaptive chain-of-thought triggering via reinforcement learning. *ArXiv*, abs/2505.11896, 2025. URL <https://api.semanticscholar.org/CorpusID:278739729>.

H. Luo, Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, Q. Lin, S. Chen, and D. Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023.

MAA. American invitational mathematics examination - AIME. American Invitational Mathematics Examination - AIME 2024, Feb. 2024. URL <https://maa.org/math-competitions/american-invitational-mathematics-examination-aime>. Accessed: 2024-05-14.

A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto. s1: Simple test-time scaling, 2025. URL <https://arxiv.org/abs/2501.19393>.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

B. Peng, E. Alcaide, Q. G. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. N. Chung, L. Derczynski, X. Du, M. Grella, K. K. GV, X. He, H. Hou, P. Kazienko, J. Kocon, J. Kong, B. Koptyra, H. Lau, J. Lin, K. S. I. Mantri, F. Mom, A. Saito, G. Song, X. Tang, J. S. Wind, S. Woźniak, Z. Zhang, Q. Zhou, J. Zhu, and R.-J. Zhu. RWKV: Reinventing RNNs for the transformer era. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=7SaXczaBpG>.

H. Puerto, T. Chubakov, X. Zhu, H. T. Madabushi, and I. Gurevych. Fine-tuning with divergent chains of thought boosts reasoning through self-correction in language models, 2024.

Y. Qu, T. Zhang, N. Garg, and A. Kumar. Recursive introspection: Teaching language model agents how to self-improve. *ArXiv*, abs/2407.18219, 2024. URL <https://api.semanticscholar.org/CorpusID:271432135>.

Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=HPuSIXJaa9>.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.

C. V. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=4FWAwZtd2n>.

Y. Sui, Y.-N. Chuang, G. Wang, J. Zhang, T. Zhang, J. Yuan, H. Liu, A. Wen, S. Zhong, H. Chen, and X. Hu. Stop overthinking: A survey on efficient reasoning for large language models. *ArXiv*, abs/2503.16419, 2025.

K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

torchtune maintainers and contributors. torchtune: Pytorch’s finetuning library, Apr. 2024. URL <https://github.com/pytorch/torchtune>.

S. Toshniwal, I. Sorokin, A. Ficek, I. Moshkov, and I. Gitman. Genselect: A generative approach to best-of-n, 2025. URL <https://arxiv.org/abs/2507.17797>.

J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins. Solving math word problems with process- and outcome-based feedback, 2022. URL <https://arxiv.org/abs/2211.14275>.

H. Wang, A. Prasad, E. Stengel-Eskin, and M. Bansal. Soft self-consistency improves language model agents, 2024a. URL <https://arxiv.org/abs/2402.13212>.

P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand, Aug. 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL <https://aclanthology.org/2024.acl-long.510/>.

X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024c. URL <https://openreview.net/forum?id=y10DM6R2r3>.

A. Wettig, K. Lo, S. Min, H. Hajishirzi, D. Chen, and L. Soldaini. Organize the web: Constructing domains enhances pre-training data curation, 2025. URL <https://arxiv.org/abs/2502.10341>.

Y. Xie, K. Kawaguchi, Y. Zhao, X. Zhao, M.-Y. Kan, J. He, and Q. Xie. Self-evaluation guided beam search for reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=Bw82hwg5Q3>.

Y. Xie, A. Goyal, W. Zheng, M.-Y. Kan, T. P. Lillicrap, K. Kawaguchi, and M. Shieh. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*, 2024.

W. Xiong, H. Zhang, N. Jiang, and T. Zhang. An implementation of generative prm. <https://github.com/RLHFlow/RLHF-Reward-Modeling>, 2024.

A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, K. Lu, M. Xue, R. Lin, T. Liu, X. Ren, and Z. Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL <https://arxiv.org/abs/2409.12122>.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. Qwen3 technical report, 2025a. URL <https://arxiv.org/abs/2505.09388>.

C. Yang, Q. Si, Y. Duan, Z. Zhu, C. Zhu, Z. Lin, L. Cao, and W. Wang. Dynamic early exit in reasoning models. *ArXiv*, abs/2504.15895, 2025b.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. R. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=5Xc1ecxO1h>.

X. Ye, F. Yin, Y. He, J. Zhang, H. Yen, T. Gao, G. Durrett, and D. Chen. Longproc: Benchmarking long-context language models on long procedural generation. In *Second Conference on Language Modeling*, 2025.

W. Zeng, Y. Huang, Q. Liu, W. Liu, K. He, Z. Ma, and J. He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025a. URL <https://arxiv.org/abs/2503.18892>.

Z. Zeng, Q. Cheng, Z. Yin, Y. Zhou, and X. Qiu. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities?, 2025b. URL <https://arxiv.org/abs/2502.12215>.

A. Zhang, Y. Chen, J. Pan, C. Zhao, A. Panda, J. Li, and H. He. Reasoning models know when they're right: Probing hidden states for self-verification. In *Second Conference on Language Modeling*, 2025a.

J. Zhang, N. Lin, L. Hou, L. Feng, and J. Li. Adaptthink: Reasoning models can learn when to think. *ArXiv*, abs/2505.13417, 2025b.

Z. Zhang, C. Zheng, Y. Wu, B. Zhang, R. Lin, B. Yu, D. Liu, J. Zhou, and J. Lin. The lessons of developing process reward models in mathematical reasoning, 2025c. URL <https://arxiv.org/abs/2501.07301>.

A More results details

A.1 SSA Results

We compared SSA performance across model size, number of candidate solution k size, subcategories, and different version of SSAs.

SSA RL performance We can see SSA has strong performance gain in most of the sub categories compared to other methods in Figure 5. In addition, we also see consistent performance gain from 0.5B, 1B, and 3B model across different LLM_{ans} sizes, and from Figure 6, we can see that LLM_{ans} has more effect over the performance gain than the SSA model size.

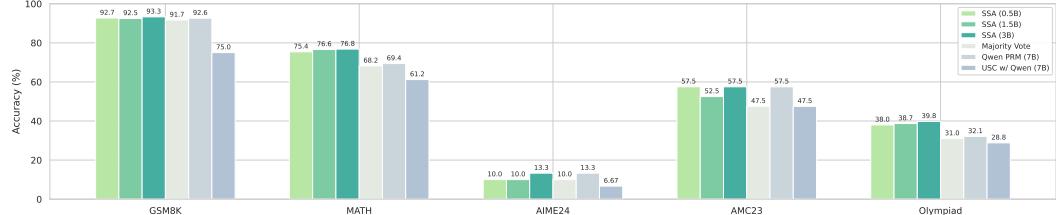


Figure 5: Compare the performance of model based on Qwen 2.5 7B with $k = 5$. SSAs are in green. We see SSA method is very effective against baseline methods.

In addition, we report all benchmark breakdown performance of SSA compared to the PRM 7B and majority vote in Table 4.

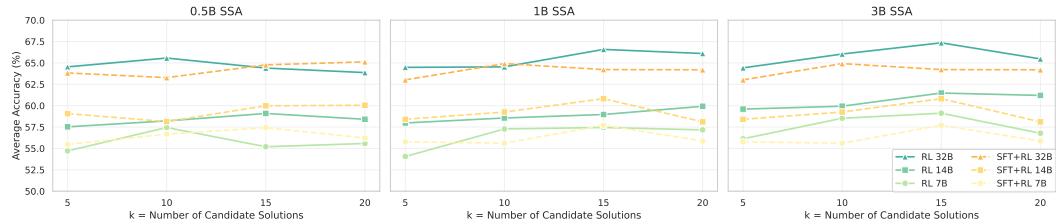


Figure 6: Compare each SSA’s performance on the average of 5 benchmarks with both RL and SFT+RL methods. SFT+RL details are in Section A.3. The same SSAs are used on top of Qwen 2.5 7B, 14B, and 32B outputs. For each model, the parallel number of candidate solutions k for SSAs are $k = 5, 10, 15, 20$.

We also report the statistical significance of all results in Table 5. For statistical testing, we employed McNemar’s test, which is appropriate for paired binary outcomes. On our combined dataset analysis (aggregating all 2563 test examples across five benchmarks), our improvements over the baseline are statistically significant ($p < 0.05$) for all configurations.

A.2 Ablation on SFT, No Think, and RL Results

We put a combined table of SFT, No Think and RL in Table 6.

A.3 Combining SFT and RL

Because we observe that the model does not give human interpretable reasoning process, we wonder whether we can improve it. Understanding the model reasoning is important for the decision making process. So we decided to use the SFT version to cold start the model behavior. We hope to have a model to give reasoning process on why they select certain answers. We will call the cold started model as SSA SFT+RL. The performance comparison is in Figure 6 and Table 7. Overall, we observe 1-2% drop of the performance with SSA SFT+RL. However, it gives us a stronger sense on the model decision making process. In addition, SFT+RL version seems to have a better performance on datasets outside of the math domain as our discussion in Section A.8

Table 4: Performance (%) on five mathematical benchmarks for 7 B, 14 B and 32 B inference models trained with RL or RL + SFT.

Benchmark		Method		7B										14B										32B																	
				RL					RL+SFT					RL					RL+SFT					RL					RL+SFT												
				5	10	15	20	5	10	15	20	5	10	15	20	5	10	15	20	5	10	15	20	5	10	15	20	5	10	15	20										
GSM8K	Pass@ κ	95.45	96.21	96.36	96.66	95.45	96.21	96.36	96.66	97.19	97.50	97.65	96.66	97.19	97.50	97.65	97.12	97.88	97.88	97.12	97.50	97.88	97.88	97.12	97.50	97.88	97.88	97.12	97.50	97.88	97.88	97.12									
	Majority vote	91.66	91.96	92.72	92.65	91.66	91.96	92.72	92.65	94.62	94.54	94.31	94.39	94.62	94.54	94.31	94.39	95.30	95.75	95.38	95.98	95.30	95.75	95.98	95.30	95.75	95.98	95.30	95.75	95.98	95.30	95.75	95.98								
	Owen PRM 7B	93.18	93.30	93.40	93.33	92.97	93.18	93.30	93.33	95.38	95.91	96.13	95.38	96.06	95.68	96.06	95.68	96.13	96.13	96.13	96.13	96.13	96.13	96.13	96.13	96.13	96.13	96.13	96.13	96.13	96.13	96.13	96.13								
	Owen PRM 72B	92.87	93.71	93.18	94.09	92.87	93.71	93.18	94.09	96.44	96.44	96.44	96.44	96.44	96.44	96.44	96.44	96.29	97.95	96.06	96.44	96.06	96.44	96.44	96.06	96.44	96.44	96.06	96.44	96.44	96.06	96.44	96.44	96.06	96.44						
	0.5B	92.65	92.87	92.49	92.65	92.57	92.87	92.80	92.95	94.69	94.69	94.62	94.77	94.62	94.77	94.62	94.77	94.62	94.77	94.64	94.77	94.64	94.77	94.64	94.77	94.64	94.77	94.64	94.77	94.64	94.77	94.64	94.77	94.64							
	1.5B	92.49	92.72	93.25	93.10	92.57	93.03	93.10	92.57	93.25	93.10	92.95	93.10	92.95	93.10	92.95	93.10	92.95	93.10	92.95	93.10	92.95	93.10	92.95	93.10	92.95	93.10	92.95	93.10	92.95	93.10	92.95	93.10								
	3B	93.25	93.18	93.63	93.33	92.65	93.33	92.65	93.33	92.95	94.77	94.92	94.77	94.92	94.77	94.92	94.77	94.92	94.62	94.62	94.62	94.62	94.62	94.62	94.62	94.62	94.62	94.62	94.62	94.62	94.62	94.62	94.62	94.62							
	Pass@ κ	78.00	81.40	83.40	85.00	78.00	81.40	83.40	85.00	82.00	84.20	85.60	85.80	82.00	84.20	85.60	85.80	81.80	84.00	84.00	85.80	86.40	81.80	84.00	84.00	85.80	86.40	81.80	84.00	84.00	85.80	86.40									
	Majority vote	68.20	69.40	71.60	71.20	68.20	69.40	71.60	71.20	74.40	75.20	75.00	75.40	74.40	75.20	75.00	75.40	75.20	75.00	75.00	75.00	75.00	75.00	75.00	75.00	75.00	75.00	75.00	75.00	75.00	75.00	75.00									
MATH	Owen PRM 7B	69.40	69.00	69.72	70.00	69.40	69.00	69.72	70.00	73.60	74.40	75.20	73.60	74.40	75.20	73.60	74.40	75.20	73.60	74.00	74.00	74.00	74.00	74.00	74.00	74.00	74.00	74.00	74.00	74.00	74.00	74.00	74.00	74.00							
	Owen PRM 72B	69.6	70.4	71.4	72.2	69.6	70.4	71.4	72.2	72.8	74.6	75	72.8	74.6	75	72.8	74.6	75	72.8	74.6	75	72.8	74.6	75	72.8	74.6	75	72.8	74.6	75	72.8	74.6									
	0.5B	75.40	76.20	76.80	78.20	76.80	78.00	78.60	79.60	80.60	81.20	80.40	80.60	81.40	81.60	80.60	81.40	81.60	82.40	82.00	81.80	82.40	82.00	81.80	82.40	82.00	81.80	82.40	82.00	81.80	82.40	82.00	81.80	82.40							
	1.5B	76.60	77.60	78.60	79.40	76.60	77.60	78.60	79.40	81.40	81.60	82.40	81.60	82.40	81.60	82.40	81.60	82.40	81.60	82.40	82.00	81.80	82.40	82.00	81.80	82.40	82.00	81.80	82.40	82.00	81.80	82.40	82.00								
	3B	76.80	78.80	79.80	79.80	77.20	77.40	79.20	79.40	82.00	82.60	82.40	82.00	82.60	82.40	82.00	82.60	82.40	82.00	82.60	82.40	82.00	82.60	82.40	82.00	82.60	82.40	82.00	82.60	82.40	82.00	82.60	82.40								
	Pass@ κ	16.67	20.00	26.67	30.00	16.67	20.00	26.67	30.00	16.67	23.33	26.67	30.00	16.67	23.33	26.67	30.00	16.67	23.33	26.67	30.00	16.67	23.33	26.67	30.00	16.67	23.33	26.67	30.00	16.67	23.33	26.67	30.00	16.67	23.33	26.67	30.00				
	Majority vote	10.00	13.33	16.67	16.67	10.00	13.33	16.67	16.67	20.00	23.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33	13.33					
	Owen PRM 7B	13.33	16.67	20.00	23.33	13.33	16.67	20.00	23.33	13.33	16.67	20.00	23.33	13.33	16.67	20.00	23.33	13.33	16.67	20.00	23.33	13.33	16.67	20.00	23.33	13.33	16.67	20.00	23.33	13.33	16.67	20.00	23.33	13.33	16.67	20.00					
	0.5B	10.00	13.33	16.67	16.67	10.00	13.33	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67							
	1.5B	10.00	13.33	16.67	16.67	10.00	13.33	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67	16.67							
	3B	13.33	20.00	16.67	13.33	13.33	13.33	13.33	10.00	13.33	13.33	13.33	10.00	13.33	13.33	13.33	10.00	13.33	13.33	13.33	10.00	13.33	13.33	13.33	10.00	13.33	13.33	13.33	10.00	13.33	13.33	13.33	10.00	13.33	13.33						
AIME24	Pass@ κ	67.50	80.00	85.00	67.50	80.00	85.00	67.50	80.00	85.00	80.00	82.50	72.50	80.00	80.00	82.50	72.50	80.00	80.00	82.50	72.50	80.00	80.00	82.50	72.50	80.00	80.00	82.50	72.50	80.00	80.00	82.50	72.50	80.00	80.00						
	Majority vote	47.50	52.50	57.50	57.50	47.50	52.50	57.50	57.50	50.00	55.00	55.00	50.00	55.00	55.00	50.00	55.00	55.00	55.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00				
	Owen PRM 7B	57.50	62.50	60.00	57.50	62.50	60.00	57.50	62.50	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00							
	0.5B	57.5	62.5	70	70	57.5	62.5	70	70	62.5	57.5	60	62.5	57.5	60	62.5	57.5	60	62.5	60	62.5	60	62.5	60	62.5	60	62.5	60	62.5	60	62.5	60	62.5	60	62.5	60	62.5				
	1.5B	52.50	55.00	57.50	55.00	60.00	60.00	57.50	55.00	57.50	55.00	57.50	55.00	57.50	55.00	57.50	55.00	57.50	55.00	57.50	55.00	57.50	55.00	57.50	55.00	57.50	55.00	57.50	55.00	57.50	55.00	57.50	55.00	57.50	55.00						
	3B	57.50	60.00	65.00	55.00	52.50	52.50	65.00	55.00	52.50	52.50	52.50	52.50	52.50	52.50	52.50	52.50	52.50	52.50	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00	60.00				
	Pass@ κ	41.10	47.18	50.15	51.48	41.10	47.18	50.15	51.48	46.29	51.48	56.23	46.29	51.48	56.23	46.29	51.48	56.23	46.29	51.48	56.23	46.29	51.48	56.23	46.29	51.48	56.23	46.29	51.48	56.23	46.29	51.48	56.23	46.29	51.48	56.23	46.29	51.48			
	Majority vote	31.01	37.16	35.31	31.01	32.79	35.16	35.31	31.01	32.05	34.42	35.01	33.83	34.12	35.36	34.12	35.36	34.12	35.36	34.12	35.36	34.12	35.36	34.12	35.36	34.12	35.36	34.12	35.36	34.12	35.36	34.12	35.36	34.12	35.36	34.12	35.36	34.12			
	Owen PRM 7B	33.68	35.46	35.31	33.68	35.46	35.31	35.76	35.31	35.76	35.88	35.97	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76	35.76			
Olympiad	Pass@ κ	59.74	64.96	68.32	69.63	59.74	64.96	68.32	69.63	62.82	67.24	68.78	70.44	62.82	67.24	68.78	70.44	66.84	70.30	72.18	73.12	66.84	70.30	72.18	73.12	66.84	70.30	72.18	73.12	66.84	70.30	72.18	73.12	66.84	70.30	72.18	73.12	66.84	70.30	72.18	73.12
	Majority vote	49.67	52.00	54.73	54.67	49.67	52.00	54.73	54.67	55.21	56.10	55.21	56.10	55.21	56.10	55.21	56.10	55																							

Table 5: Statistical comparison of SSA models vs. baselines on mathematical benchmarks.

Baseline	Method	7B						14B						32B					
		5		10		20		5		10		20		5		10		20	
		p-val	sig																
Majority Vote	0.5B vs. Maj	1.1e-14	*	3.5e-14	*	1.9e-05	*	7.0e-11	*	1.4e-06	*	4.5e-06	*	2.5e-16	*	3.6e-10	*	2.5e-06	*
	1B vs. Maj	4.2e-17	*	2.0e-18	*	9.6e-14	*	1.2e-12	*	1.3e-09	*	7.4e-12	*	1.4e-18	*	1.8e-17	*	1.6e-16	*
	3B vs. Maj	5.8e-25	*	6.3e-23	*	5.1e-15	*	3.1e-17	*	5.1e-14	*	4.3e-14	*	2.0e-20	*	3.0e-15	*	8.4e-15	*
Owen PRM	0.5B vs. PRM	6.7e-06	*	1.0e-06	*	4.6e-04	*	1.1e-08	*	1.8e-04	*	1.4e-03	*	3.3e-13	*	2.8e-11	*	6.6e-07	*
	1B vs. PRM	1.1e-06	*	1.2e-06	*	2.7e-08	*	9.3e-10	*	8.1e-06	*	9.0e-06	*	1.0e-13	*	1.5e-15	*	2.7e-13	*
	3B vs. PRM	1.0e-10	*	1.4e-08	*	4.2e-10	*	1.5e-13	*	2.0e-08	*	6.9e-08	*	5.5e-15	*	3.5e-15	*	1.7e-12	*

Note: All results are on the combined of all 5 benchmarks. The * indicates statistical significance ($p < 0.05$). All SSA models (0.5B, 1B, 3B) show significant improvement over both Majority vote and PRM baselines across all inference model sizes (7B, 14B, 32B).

Table 6: Ablation of different training methods. Including SFT method, No Thinking Method, and RL methods trained with LLM_{ans} $k = 5$ on GSM8K train data only. We report LLM_{ans} $k = 5, 10$ results as accuracy (%)

Model	GSM8K		MATH		AIME24		AMC23		Olympiad		Average	
	k=5	k=10	k=5	k=10	k=5	k=10	k=5	k=10	k=5	k=10	k=5	k=10
<i>Baseline</i>												
Majority Vote	91.66	91.96	68.20	69.40	10	13.33	47.50	52.50	31.01	32.79	49.67	52.0
PRM	92.57	93.18	69.4	69.00	13.33	16.67	57.5	62.5	32.05	34.42	52.97	55.15
<i>0.5B</i>												
SFT	91.51	92.04	66.0	66.6	10.0	10.0	57.5	50.0	30.86	33.38	51.17	50.4
RL No-Think	91.43	91.58	62.4	61.2	10.0	10.0	52.5	47.5	28.49	28.19	48.96	47.69
RL	92.42	93.1	58.6	57.2	10.0	13.33	60.0	50.0	28.04	27.89	49.81	48.3
<i>1.5B</i>												
SFT	91.51	91.66	72.4	70.4	10.0	10.0	50.0	42.5	36.05	30.56	51.99	49.02
RL No-Think	92.65	93.1	63.2	63.2	10.0	16.67	52.5	52.5	31.75	32.94	50.02	51.68
RL	92.65	93.1	71.6	73.0	10.0	20.0	47.5	55.0	34.57	35.91	51.26	55.4
<i>3B</i>												
SFT	91.13	91.58	68.4	66.6	6.67	10.0	45.0	40.0	29.38	31.31	48.12	47.9
RL No-Think	92.8	93.25	72.6	72.8	10.0	20.0	57.5	52.5	36.8	38.28	53.94	55.37
RL	93.18	93.1	75.0	74.8	13.33	16.67	60.0	52.5	34.57	37.69	55.22	54.95

Table 7: Average Performance (%) of SSA RL and SSA SFT+RL over five benchmarks. The answers are generated with Qwen 32B models. The second row indicates the number of sampled answers (k).

Method	RL				SFT + RL			
	5	10	15	20	5	10	15	20
Pass@k	66.84	70.30	72.18	73.12	66.84	70.30	72.18	73.12
Majority vote	59.51	62.63	61.63	61.77	59.51	62.63	61.63	61.77
Qwen PRM	59.45	59.54	59.80	59.47	59.45	59.54	59.80	59.47
SSA (0.5B)	64.54	65.57	64.40	63.87	63.83	63.28	64.77	65.13
SSA (1.5B)	64.48	64.54	66.58	66.10	63.02	64.92	64.22	64.20
SSA (3B)	64.42	66.04	67.35	65.47	63.71	64.80	65.73	65.14

A.4 Increasing k during training

We see that the inference k could lead to the potential improvements of the performance. Would the same hold if we improve the k during training. For the original design, we use $k = 5$ for the training. For comparison, we train the model with $k = 8$ to see its performance. The results are presented in Figure 7. We see that training longer context does not help with the performance. In fact it has lower performance on average for the dataset. It might due to longer context creates more same answers, and it would make the model to choose more depends on the majority vote than distinguish the differences.

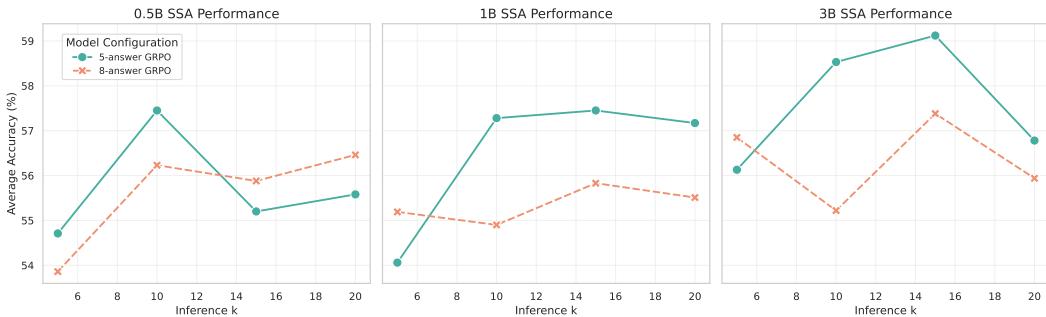


Figure 7: Compare training k and their effect. It seems that increasing context length to longer k during training does not out perform the shorter context.

Table 8: Aggregation overhead per question

Answer K	32	64	128
Owen PRM (7B)	5.25	5.41	5.80
SSA RL (3B)	9.28	7.94	7.02

A.5 Scaling K during inference and Two-stage SSA

To handle large K under context limits, we use a simple two-stage adaptation of SSA.

Stage 1. We form ℓ_2 overlapping groups by taking *evenly spaced, cyclic windows* of length ℓ_1 over the K candidates. Concretely, with start indices $s_i = \left\lfloor \frac{iK}{\ell_2} \right\rfloor$ for $i = 0, \dots, \ell_2 - 1$, the i -th group is $G_i = [a_{(s_i+t) \bmod K}]_{t=0}^{\ell_1-1}$. We run SSA independently on each G_i to produce one intermediate winner.

Stage 2. We concatenate the ℓ_2 intermediate winners and run SSA once more to produce the final answer.

This makes ℓ_2+1 SSA calls: ℓ_2 calls on inputs of size ℓ_1 , then one call on ℓ_2 . We use $\ell_1=15$ in all experiments; thus $\ell_2=\lceil K/\ell_1 \rceil$ gives $\ell_2=3$ for $K=32$, $\ell_2=5$ for $K=64$, and $\ell_2=9$ for $K=128$. By construction, groups *overlap* whenever $\ell_1 > K/\ell_2$ (our default), yielding a replication factor $(\ell_1\ell_2)/K$ that adds controlled redundancy for robustness at modest extra compute. The details are in Algorithm 1.

In addition to the performance reported in the Table 3, we also report the aggregation overhead over the AMC23 dataset. The results are in Table 8.

Algorithm 1 Two-stage SSA

Require: Question x ; K candidates $A = \{a_0, \dots, a_{K-1}\}$; group size ℓ_1 ; group count ℓ_2

```

1: if  $K \leq \ell_1$  then
2:   return SSA( $x, A$ )
3: end if
4: stage1  $\leftarrow []$  ▷ will hold  $\ell_2$  winners (full CoTs)
5: for  $i = 0$  to  $\ell_2 - 1$  do ▷ Stage 1:  $\ell_2$  calls, each on  $\ell_1$  answers
6:    $s_i \leftarrow \left\lfloor \frac{i \cdot K}{\ell_2} \right\rfloor$  evenly spaced starts
7:    $G_i \leftarrow [a_{((s_i+t) \bmod K)} : t = 0, \dots, \ell_1 - 1]$  cyclic window (overlap-allowed)
8:    $\hat{y}_i \leftarrow \text{SSA}(x, G_i)$  infer the group's final answer
9:    $a_{G_i} \leftarrow \text{MapBack}(\hat{y}_i, G_i)$  pick the chosen candidate's full CoT
10:  stage1.append( $a_{G_i}$ )
11: end for ▷ Stage 2: aggregate  $\ell_2$  winners
12: return SSA( $x, \text{stage1}$ )

```

A.6 Error Analysis

We conduct error analysis of the SSA outputs. We use SSA 3B model with five answer samples. Across all datasets the vast majority of correct cases are simply *copied* from a candidate that already contains the ground-truth answer. Hence the main value of SSA lies in finding the right candidate. When the ground-truth answer is absent from the sample set, SSA fails in two ways: it either chooses the majority wrong answer ($y^* \notin C$ *copied*) or try to come up a new but still wrong answer ($y^* \notin C$ *synthesized*). While SSA sometimes come up with the correct answer even the ground truth are not provided in the inference, when we manually examine the answers in this category most are in wrong format that extraction failed. It confirms an input quality bottleneck: if no correct candidate exists, the policy has difficulty to recover it. Conversely, errors with a correct candidate present ($y^* \in C$ / *copied wrong* or $y^* \in C$ / *synthesized wrong*) are much smaller, showing that SSA rarely mis-ranks truly correct answers. In order to understand whether SSA can synthesize new answers, we designed addition experiments to cut off last 10% of the answers, and our results are in Appendix A.7. It

Table 9: Accuracy and error composition for the 3 B SSA ($k=5$). Columns are percentages of the whole dataset. $y^* \in C$ means ground-truth answer y^* included among the 5 candidates C .

Acc.↑	Correct			Wrong		
	$y^* \in C$ copied	$y^* \notin C$ synthesized	$y^* \notin C$ copied	$y^* \notin C$ synthesized	$y^* \in C$ copied	$y^* \in C$ synthesized
GSM8K	93.3	87.7	5.5	3.3	1.0	2.3
MATH	79.2	75.6	3.6	9.4	3.6	6.6
AMC23	57.5	52.5	5.0	12.5	7.5	12.5
AIME24	13.3	13.3	0.0	23.3	50.0	3.3
Olympiad	39.8	34.9	4.9	27.6	24.6	5.6

shows that the SSA has synthesize ability when answers are all removed, and the accuracy has some degradation.

A.7 Copy or Synthesize?

Our error analysis reveals that the majority of correct answers are copied from the provided responses. To test whether the SSA can synthesize answers, we created a variant of the dataset where the last 10% of every answer is cut off. This ensures that the final answers are not directly available in the input, requiring the model to synthesize them.

Table 10: Performance comparison between original responses and those with last 10% removed.

Method	Datasets					Avg
	GSM8K	MATH	AIME24	AMC23	Olympiad	
<i>Original (complete answers)</i>						
SSA	93.25	76.80	13.33	57.50	39.76	56.13
<i>Truncated (last 10% removed)</i>						
SSA + RL	74.22	52.80	0.00	22.50	23.00	34.50
SSA + RL + SFT	89.99	67.80	6.67	40.00	31.01	47.09

As shown in Table 10, performance significantly drops when answers must be synthesized rather than copied. The RL method alone achieves only 34.50% average accuracy, while combining RL with SFT improves this to 47.09%. Notably, the RL+SFT approach recovers much of the performance gap, particularly on GSM8K (89.99% vs. original 93.25%), suggesting effective synthesis capabilities for easy problems. However, since pure copy would only result 0%, this suggests that SSA is not just copy but able to synthesize correct results.

A.8 Generalization to other tasks

Table 11 reports accuracy on three general tasks benchmarks. On **ARC-C** and **MMLU-Pro** the SFT + RL SSA mostly beats the majority-vote baseline [Clark et al., 2018, Wang et al., 2024c]. Pure RL SSAs give smaller (sometimes negative) gains, suggesting that an SFT warm-start improves out-of-domain transfer. Pure SFT SSAs exhibit varied performance: strong at 3B scale but significantly weaker at smaller scales (0.5B and 1.5B), suggesting SFT’s generalization capability strongly depends on model capacity for out domain generalization. On **TruthfulQA** (truthfulness/adversarial) none of the SSAs can consistently outperform majority voting, suggesting that truthfulness might not be beneficial from answer selection. Tackling truthfulness might require different goal than reasoning among answers.

Table 11: Model evaluation results on ARC, MMLU-PRO, TruthfulQA benchmarks

Metric	Qwen2.5-7B-Instruct		SSA RL			SSA SFT + RL			SSA SFT		
	pass@1	Majority Vote (5)	0.5B	1.5B	3B	0.5B	1.5B	3B	0.5B	1.5B	3B
ARC-C	88.65	91.38	88.05	91.3	88.99	91.72	91.38	91.81	69.88	90.02	92.06
MMLU-PRO	43.63	49.24	33.19	43.09	39.84	46.28	48.01	50.37	25.76	43.93	51.41
TruthfulQA	62.43	66.23	66.52	67.11	64.62	66.08	64.47	63.89	49.85	63.89	66.23

Table 12: Efficiency Evaluation. We report end-to-end runtime on one RTX 6000 Ada.

Method	Base-LM passes	Aggregator time	Total time (s) ↓	Overall Accuracy (%) ↑
Qwen2.5-7B	1	—	556.12	45.5
Majority vote ($k=5$)	$5 \times$ Qwen2.5-7B	—	2 780.6	49.67
Qwen PRM 7B ($k=5$)	$5 \times$ Qwen2.5-7B	21.9	$2 780.6 + 21.9 = 2 802.5$	52.97
SSA RL 3B ($k=5$, ours)	$5 \times$ Qwen2.5-7B	25.7	$2 780.6 + 25.7 = 2 806.3$	56.13
SimpleRL 7B (seq. RL)	1 pass	—	3 273.2	58.56

“Base-LM passes” = number of forward decodes of Qwen-2.5-7B-Instruct (556.12 s per pass). Aggregator time is measured separately.

A.9 Efficiency Evaluation

We now evaluate the amount of compute required for each aggregation strategy. Compared to other methods which considers each sample independently, our method considers them jointly, generating longer sequence which can incur high computation cost.

Compute Cost (Wall-clock) For more details, we conduct an actual measure based on the real settings. The results are in Table 12. All measurements were taken on a single NVIDIA RTX 6000 Ada with the AMC23 benchmark (40 questions). In practice, parallelized sampling from the LLM can significantly speed up the time required.

Inference Cost To measure raw compute (flops), we follow the formula from Kaplan et al. [2020]. It uses roughly $C_{forward} \approx 2N$, and the backward pass is about 2 times which is $4N$. So the total is $C \approx 6N$ FLOPs per training token, with N as the model parameter and C as the non embedding training compute. We use D as token length.

For inference cost we will use $C_{forward} \approx 2ND$ per query. For $k=5$, input context $D \approx 5 * 1000$. So the SSA method needs a 7B base model with inference cost $2ND = 2 * 7 * 10^9 * (1000 * 5) = 70TFLOPs$. SSA itself has a constant compute overhead of $2ND = 2 * 3 * 10^9 * (5000 + 60) \approx 30TFLOPs$ since SSA’s compute cost depends on the input token length not the inference model size. The total for SSA would be $70 + 30 = 100TFLOPs$. In comparison, the sequential RL approach would result $2ND = 2 * 7 * 10^9 * 8000 \approx 110TFLOPs$. This method scales favourably to larger bases (SSA 350 TFLOPs vs Sequential RL 512 TFLOPs on a 32B model).

A.10 Stability Analyses

A.10.1 Effect of Candidate Ordering

A potential concern for methods that concatenate K candidates is that the relative order of those candidates might influence the SSA’s decision. To test this, we evaluate the 3B SSA (RL) under three independent random permutations of the K responses of 7B answer model at inference time, holding everything else fixed. As shown in Table 13, we do not observe meaningful differences when changing the order.

A.10.2 Training different random seed

Reinforcement learning can exhibit sensitivity to initialization. We therefore train SSA 3B (RL) with another seed 16 using identical data and hyperparameters, and evaluate the resulting checkpoints under the same protocol as the main results of 7B answer model. Table 14 reports per- K accuracy. We do not see significant differences.

A.11 Training Data Statistics

Table 15 summarizes the distribution of instance-level correctness in the raw training pool prior to filtering. For each question we sampled $K=5$ candidate solutions using top- k sampling. “ $m/5$ ” indicates that exactly m of the five candidates match the reference answer.

Filtering and preprocessing. We apply two light filters before RL training:

Table 13: **Ordering sensitivity.** SSA 3B (RL) evaluated with the original concatenation order vs. three random permutations at inference time. Values are accuracy (%).

Metric	5	10	15	20
Original (%)	56.13	58.53	59.12	56.78
Avg. over 3 random permutations (%)	56.06	58.70	59.40	58.47

Table 14: **Training-seed sensitivity.** SSA 3B (RL) trained with different random seeds. Values are accuracy (%).

Metric	5	10	15	20
Original Random Seed 42(%)	56.13	58.53	59.12	56.78
Random Seed 16 (%)	56.09	57.27	59.49	56.64

Table 15: **Correctness distribution** in the raw training pool (five sampled responses per question). Counts and column percentages are shown.

Correct	GSM8K	MATH	Combined
0/5	618 (8.3%)	3170 (26.4%)	3788 (19.5%)
1/5	218 (2.9%)	1007 (8.4%)	1225 (6.3%)
2/5	198 (2.6%)	841 (7.0%)	1039 (5.3%)
3/5	268 (3.6%)	881 (7.3%)	1149 (5.9%)
4/5	477 (6.4%)	1157 (9.6%)	1634 (8.4%)
5/5	5694 (76.2%)	4944 (41.2%)	10638 (54.6%)
Total	7473 (100.0%)	12000 (100.0%)	19473 (100.0%)

- **Validity filter.** We discard instances where more than one of the five samples is NULL (unparseable/empty). This removes $\approx 0.4\%$ of GSM8K and $\approx 9.8\%$ of MATH instances.
- **Length filter.** Answer sets exhibit a long-tailed length distribution; some concatenations exceed 8k tokens due to looping or unbounded reasoning. To control VRAM and remove pathological traces, we drop instances whose concatenated prompt+answers exceed 4k tokens. This reduces the pool from $\sim 19k$ to $\sim 17k$ instances and slightly denoises the supervision.

These filters are minimal (no step-level labeling) and aimed purely at stabilizing training; we did not tune them for accuracy.

B Prompt Details

For SSA method, we trained and evaluate it with the following prompt:

SSA Prompt: A conversation between User and Assistant. The user provide a question and some proposed answers. The Assistant first evaluate each answers individually, check whether each answer directly addresses the original question, assess the correctness of each answer based on logical reasoning, calculations, and accuracy relative to the question. After thorough evaluation, identify one correct answer. If the correct answer is not in the provided proposed answers, the Assistant will combine the correct answer with the proposed answers and provide the correct answer. The reasoning process and answer are enclosed within `<think></think>` and `<answer></answer>` tags, respectively, i.e., `<think>reasoning process here</think> <answer>answer here</answer>`.

Figure 8: Example prompt for SSA. For reward extraction, we will use rule based extraction to extract anything inside `<think></think>` and `<answer></answer>`. If the output matches the structure and able to extract some values we will provide minimal format reward.

SSA No-Think Prompt: A conversation between User and Assistant. The user provide a question and some proposed answers. The Assistant answer the question based on the proposed answers. The answer is enclosed within `<answer></answer>` tag, i.e., `<answer>answer here</answer>`.

Figure 9: Example prompt for SSA

USC Prompt: You are a helpful assistant. The user provide a question and some proposed answers. The Assistant first evaluate each answers individually, check whether each answer directly addresses the original question, assess the correctness of each answer based on logical reasoning, calculations, and accuracy relative to the question. After thorough evaluation, identify one correct answer based on majority consensus. The reasoning process and answer are enclosed within `<think></think>` and `<answer></answer>` tags, respectively, i.e., `<think>reasoning process here</think> <answer>answer here</answer>`.

Figure 10: Example prompt for USC prompt to the Instruct model

SFT Data Construction Prompt: Here is a question and some proposed answers. You need to evaluate each answers individually, check whether each answer directly addresses the original question, assess the correctness of each answer based on logical reasoning, calculations, and accuracy relative to the question. After thorough evaluation, identify one correct answer. If the correct answer is not in the provided proposed answers, the Assistant will combine the correct partial responses to proposed answers and provide the correct answer. Make the reasoning process concise and to the point. The reasoning process and answer are enclosed within `<think></think>` and `<answer></answer>` tags, respectively, i.e., `<think>reasoning process here</think> <answer>answer here</answer>`. Here is the Question and proposed answers: **Question and 5 Answers**.

Here is the labeled answer with answer tags '`<answer>Ground Truth Answer</answer>`', and most likely the correct answer. If none of the proposed answers is correct, come up with the reasoning process to arrive at the actual correct final answer. **Attention: Do not ever mention the answer is given in the response, but you need to justify it from the problem and proposed answers!**

Figure 11: Example prompt for constructing SFT dataset

C Example Results

A conversation between User and Assistant. The user provide a question and some proposed answers. The Assistant first evaluate each answers individually, check whether each answer directly addresses the original question, assess the correctness of each answer based on logical reasoning, calculations, and accuracy relative to the question. After thorough evaluation, identify one correct answer. If the correct answer is not in the provided proposed answers, the Assistant will combine the correct answer with the proposed answers and provide the correct answer. The reasoning process and answer are enclosed within `<think></think>` and `<answer></answer>` tags, respectively, i.e., `<think>reasoning process here</think> <answer>answer here</answer>`.

User: Question: What is the value of $\sum_{k=1}^9 (2^k - 1)^3 + 4^k - 3^k + 6^k - 5^k + \dots + 18^k - 17^k$? Below are the proposed answers:

Answer 1: To solve the given expression $\sum_{k=1}^9 (2^k - 1)^3 + 4^k - 3^k + 6^k - 5^k + \dots + 18^k - 17^k$, we can observe that it consists of pairs of terms where each pair is of the form $(2k)^3 - (2k-1)^3$. Let's simplify the general term in the series: $\sum_{k=1}^9 (2k)^3 - (2k-1)^3$. Using the binomial expansion for cubes, we have: $(2a)^3 - b^3 = (a-b)(a^2 + ab + b^2)$, where $a = 2k$ and $b = 2k-1$. Substituting these values, we get: $\sum_{k=1}^9 (2k)^3 - (2k-1)^3 = \sum_{k=1}^9 [(2k)^2 + (2k-1)^2](2k-1)$. Now, let's expand the quadratic expression: $\sum_{k=1}^9 [(2k)^2 + (2k-1)^2] = \sum_{k=1}^9 [4k^2 - 2k + 1] = \sum_{k=1}^9 [4k^2 - 4k + 1]$. Adding these together: $\sum_{k=1}^9 [4k^2 - 4k + 1] = 12k^2 - 4k + 1$. This is the sequence goes up to 18, which corresponds to $\sum_{k=1}^9 (2k)^2 = 12k^2 - 6k + 1$. Next, we can split this into three separate sums: $\sum_{k=1}^9 6k = 6 \cdot \frac{9(9+1)}{2} = 270$. Multiplying by 12: $12 \cdot 270 = 3240$. Calculating each sum separately: $\sum_{k=1}^9 1 = 9$. Sum of squares: $\sum_{k=1}^9 k^2 = \frac{9(9+1)(2 \cdot 9 + 1)}{6} = 285$. Sum of integers: $\sum_{k=1}^9 k = \frac{9(9+1)}{2} = 45$. Substituting these values back into our expression: $3240 - 285 - 45 = 2905$.

Answer 2: To find the value of the expression $\sum_{k=1}^9 (2^k - 1)^3 + 4^k - 3^k + 6^k - 5^k + \dots + 18^k - 17^k$, we can observe that it is a sum of differences of cubes in pairs. Let's break it down step by step. First, consider the general term in the sequence: $\sum_{k=1}^9 (2^k - 1)^3$. For $k = 1, 2, 3, \dots, 9$, we need to simplify the expression $(2^k - 1)^3$. Using the identity for the difference of cubes, $(a^3 - b^3) = (a-b)(a^2 + ab + b^2)$, where $a = 2^k$ and $b = 2k-1$, we get: $(2^k - 1)^3 = (2k - 2k+1)(2^k)^2 + (2k-1)^2(2k-1)$. Simplifying inside the parentheses: $(2^k)^2 = 4k^2$ and $(2k-1)^2 = 4k^2 - 4k + 1$. Adding these together: $4k^2 + 4k^2 - 4k + 1 = 8k^2 - 4k + 1$. Now, we need to sum this expression from $k=1$ to $k=9$ (since the sequence goes up to 18, which corresponds to $k=9$). We can split this into three separate sums: $\sum_{k=1}^9 6k = 6 \cdot \frac{9(9+1)}{2} = 270$. Sum of squares: $\sum_{k=1}^9 k^2 = \frac{9(9+1)(2 \cdot 9 + 1)}{6} = 285$. Sum of integers: $\sum_{k=1}^9 k = \frac{9(9+1)}{2} = 45$. Substituting these values back into our expression: $270 - 285 - 45 = -60$.

Answer 3: To find the value of the expression $\sum_{k=1}^9 (2^k - 1)^3 + 4^k - 3^k + 6^k - 5^k + \dots + 18^k - 17^k$, we can observe that it consists of pairs of terms where each pair is of the form $(2n)^3 - (2n-1)^3$. First, let's simplify the general term in the sequence: $\sum_{n=1}^9 (2n)^3 - (2n-1)^3$. Using the binomial expansion for cubes, we have: $(2a)^3 - b^3 = (a-b)(a^2 + ab + b^2)$, where $a = 2n$ and $b = 2n-1$. So, $\sum_{n=1}^9 (2n)^3 - (2n-1)^3 = \sum_{n=1}^9 [(2n)^2 + (2n-1)^2](2n-1)$. Now, let's expand the quadratic expression: $\sum_{n=1}^9 [(2n)^2 + (2n-1)^2] = \sum_{n=1}^9 [4n^2 - 4n + 1] = \sum_{n=1}^9 [4n^2 - 4n + 1]$. Now, we need to sum this expression from $n=1$ to $n=9$. We can split this into three separate sums: $\sum_{n=1}^9 6n = 6 \cdot \frac{9(9+1)}{2} = 270$. Sum of squares: $\sum_{n=1}^9 n^2 = \frac{9(9+1)(2 \cdot 9 + 1)}{6} = 285$. Sum of integers: $\sum_{n=1}^9 n = \frac{9(9+1)}{2} = 45$. Substituting these values back into our expression: $270 - 285 - 45 = -60$.

Answer 4: To find the value of the expression $\sum_{k=1}^9 (2^k - 1)^3 + 4^k - 3^k + 6^k - 5^k + \dots + 18^k - 17^k$, we can observe that it is a sum of differences of cubes in pairs. Specifically, it can be written as: $\sum_{k=1}^9 (2^k - 1)^3 + (4^k - 3^k) + (6^k - 5^k) + \dots + (18^k - 17^k)$. We can use the identity for the difference of cubes, which states $(a^3 - b^3) = (a-b)(a^2 + ab + b^2)$. For each pair $(2^k - 1)^3 - (2k-1)^3 = (2k - 2k+1)(2^k)^2 + (2k-1)^2(2k-1)$. The expression simplifies to: $12k^2 - 4k + 1$. Now, we need to sum these simplified terms from $k=1$ to $k=9$. We can split this into three separate sums: $\sum_{k=1}^9 6k = 6 \cdot \frac{9(9+1)}{2} = 270$. Sum of squares: $\sum_{k=1}^9 k^2 = \frac{9(9+1)(2 \cdot 9 + 1)}{6} = 285$. Sum of integers: $\sum_{k=1}^9 k = \frac{9(9+1)}{2} = 45$. Substituting these values back into our expression: $270 - 285 - 45 = -60$.

Answer 5: To solve the expression $\sum_{k=1}^9 (2^k - 1)^3 + 4^k - 3^k + 6^k - 5^k + \dots + 18^k - 17^k$, we can observe that it consists of pairs of terms where each pair is of the form $(2n)^3 - (2n-1)^3$. First, let's simplify the general term in the sequence: $\sum_{n=1}^9 (2n)^3 - (2n-1)^3$. Using the binomial expansion for cubes, we have: $(2a)^3 - b^3 = (a-b)(a^2 + ab + b^2)$. For each pair $(2n)^3 - (2n-1)^3 = (2n - 2n+1)(2n)^2 + (2n-1)^2(2n-1)$. The expression simplifies to: $12n^2 - 6n + 1$. Now, we need to sum this expression from $n=1$ to $n=9$ (since the sequence goes up to 18 and 17, which correspond to $n=1$ to $n=9$). We can split this into three separate sums: $\sum_{n=1}^9 6n = 6 \cdot \frac{9(9+1)}{2} = 270$. Sum of squares: $\sum_{n=1}^9 n^2 = \frac{9(9+1)(2 \cdot 9 + 1)}{6} = 285$. Sum of integers: $\sum_{n=1}^9 n = \frac{9(9+1)}{2} = 45$. Substituting these values back in, we get: $270 - 285 - 45 = -60$.

Assistant: `<think>`Each proposed answer simplifies the general term and sums the series. The key is to recognize that the series is composed of pairs of cubes, and each pair simplifies to $12n^2 - 6n + 1$. The sums of squares and integers from 1 to 9 are well-known and can be calculated directly. The total sum is the sum of these three parts: $12 \cdot 270 - 6 \cdot 45 + 9$. Calculating each part: $12 \cdot 270 = 3240$, $6 \cdot 45 = 270$, and $9 = 9$. Summing these: $3240 - 270 + 9 = 3159$. All answers arrive at the same total, confirming the correctness of the calculation.`</think>` `<answer>3159</answer>`

Figure 12: Example output when we cut off the last 10% of all answers to test whether SSA can synthesize. SSA SFT+RL 3B can synthesize and continue to get the final correct answer from AMC23 data. We can see the final answer 3159 does not appear anywhere in the provided solutions.

A conversation between User and Assistant. The user provide a question and some proposed answers. The Assistant first evaluate each answers individually, check whether each answer directly addresses the original question, assess the correctness of each answer based on logical reasoning, calculations, and accuracy relative to the question. After thorough evaluation, identify one correct answer. If the correct answer is not in the provided proposed answers, the Assistant will combine the correct answer with the proposed answers and provide the correct answer. The reasoning process and answer are enclosed within `<think></think>` and `<answer></answer>` tags, respectively, i.e., `<think>reasoning process here</think><answer>answer here</answer>`. User: Question: A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?
Below are the proposed answers:

Answer 1:
To determine the total number of bolts needed for the robe, we start by identifying the amount of each type of fiber required:
- The robe requires 2 bolts of blue fiber.
- It requires half as much white fiber as blue fiber, which means it needs $\frac{1}{2}$ bolt of white fiber.
Adding these together gives us the total number of bolts:

$$\frac{1}{2} + 2 = 2.5$$

Answer 2:
To determine the total number of bolts needed for the robe, we start by identifying the amount of each type of fiber required:
1. The robe requires 2 bolts of blue fiber.
2. The amount of white fiber required is half the amount of blue fiber, which is $\frac{1}{2}$ bolt.
Adding these together gives the total number of bolts needed:

$$2 + \frac{1}{2} = 2.5$$

Answer 3:
To determine the total number of bolts of fiber needed for the robe, we start by identifying the amount of each type of fiber required:
1. The robe requires 2 bolts of blue fiber.
2. The amount of white fiber required is half that of the blue fiber. Therefore, the amount of white fiber needed is $\frac{1}{2}$ bolt.
Next, we add the bolts of blue fiber and white fiber together to find the total number of bolts needed:

$$2 + \frac{1}{2} = 2.5$$

Answer 4:
A robe takes 2 bolts of blue fiber and half that amount in white fiber, which means it takes 1 bolt of white fiber (since half of 2 is 1). Therefore, the total number of bolts needed is $2 + 1 = 3$ bolts.

Answer 5:
A robe takes 2 bolts of blue fiber and half that amount in white fiber, which means it takes 1 bolt of white fiber (since half of 2 is 1). In total, the robe takes 2 bolts of blue fiber + 1 bolt of white fiber = 3 bolts.

Assistant: `<think>`Each proposed answer correctly identifies that the robe requires 2 bolts of blue fiber and half that amount in white fiber, which is 1 bolt. The total bolts needed are then $2 + 1 = 3$ bolts. All answers arrive at the same conclusion through correct reasoning and calculations. Since all answers are consistent and correct, the most straightforward and clear answer is the one that explicitly states the total as 3 bolts, which is Answer 5. Therefore, the correct answer is 3 bolts.`</think> <answer>3</answer>`

Figure 13: SSA SFT Example. We can see there is a more clear reasoning on the response on the decision process.

A conversation between User and Assistant. The user provide a question and some proposed answers. The Assistant first evaluate each answers individually, check whether each answer directly addresses the original question, assess the correctness of each answer based on logical reasoning, calculations, and accuracy relative to the question. After thorough evaluation, identify one correct answer. If the correct answer is not in the provided proposed answers, the Assistant will combine the correct answer with the proposed answers and provide the correct answer. The reasoning process and answer are enclosed within `<think></think>` and `<answer></answer>` tags, respectively, i.e., `<think>reasoning process here</think>` `<answer>answer here</answer>`.

User: Question: Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did he make?
Below are the proposed answers:

Answer 1:
First, we need to calculate the new value of the house after the repairs. The original cost of the house was \$80,000, and Josh spent an additional \$50,000 on repairs. Therefore, the total cost before considering the increase in value is: $80,000 + 50,000 = 130,000$. The repairs increased the value of the house by 150%. To find the new value of the house, we need to calculate 150% of the original purchase price (\$80,000) and add it to the original purchase price. $150\% \text{ of } 80,000 = 1.5 \times 80,000 = 120,000$. Adding this to the original purchase price gives us the new value of the house: $80,000 + 120,000 = 200,000$. Now, we subtract the total amount Josh spent (the initial purchase price plus the repair costs) from the new value of the house to determine his profit: $200,000 - 130,000 = 70,000$. Therefore, Josh made a profit of \$70,000.

Answer 2:
First, we need to calculate the new value of the house after the repairs. The original cost of the house was \$80,000, and Josh spent an additional \$50,000 on repairs. Therefore, the total cost before considering the increase in value is: $80,000 + 50,000 = 130,000$. Next, we know that the repairs increased the value of the house by 150%. An increase of 150% means the value of the house became 250% (or 2.5 times) its original value before the repairs. So, the new value of the house is: $130,000 \times 2.5 = 325,000$. Now, to find the profit, we subtract the total cost from the new value of the house: $325,000 - 130,000 = 195,000$. Therefore, the profit Josh made is: \$195,000.

Answer 3:
First, let's calculate the new value of the house after the repairs. The original cost of the house was \$80,000, and Josh spent an additional \$50,000 on repairs. Therefore, the total cost before considering the increase in value is: $80,000 + 50,000 = 130,000$. Next, we need to determine the new value of the house after the 150% increase in value. A 150% increase means the value of the house becomes 250% (100% original value + 150% increase) of its pre-repair value. The pre-repair value of the house is \$130,000, so the new value is: $130,000 \times 2.5 = 325,000$. Now, to find the profit, we subtract the total cost from the new value of the house: $325,000 - 130,000 = 195,000$. Therefore, the profit Josh made is: \$195,000.

Answer 4:
First, let's calculate the new value of the house after the repairs. The original purchase price of the house was \$80,000, and Josh spent an additional \$50,000 on repairs. Therefore, the total cost before considering the increase in value is: $80,000 + 50,000 = 130,000$. Next, we need to determine the new value of the house after it increased by 150%. An increase of 150% means the value of the house becomes 250% (or 2.5 times) its original purchase price. So, the new value of the house is: $80,000 \times 2.5 = 200,000$. Now, we can find out how much profit Josh made by subtracting his total cost from the new value of the house: $200,000 - 130,000 = 70,000$. Therefore, the profit Josh made is: \$70,000.

Answer 5:
First, we need to determine the new value of the house after the repairs. The original cost of the house was \$80,000, and Josh spent an additional \$50,000 on repairs. Therefore, the total cost before any increase in value is: $80,000 + 50,000 = 130,000$. The problem states that the value of the house increased by 150% due to the repairs. An increase of 150% means the value of the house became 250% (or 2.5 times) its original cost after repairs. So, the new value of the house is: $130,000 \times 2.5 = 325,000$. To find the profit, we subtract the total cost from the new value of the house: $325,000 - 130,000 = 195,000$. Therefore, the profit Josh made is: \$195,000.

Assistant: `<think>reasoning process here</think>` `<answer>195000</answer>`,

Figure 14: SSA RL Example. We can see the response reasoning process is just a copy of the instruction template. There are no meaningful content there

A conversation between User and Assistant. The user provide a question and some proposed answers. The Assistant answer the question based on the proposed answers. The answer is enclosed within `<answer></answer>` tag, i.e., `<answer>answer here</answer>`.

User: Question: Convert the point $(0,3)$ in rectangular coordinates to polar coordinates. Enter your answer in the form (r,θ) , where $r > 0$ and $0 \leq \theta < 2\pi$. Below are the proposed answers:

Answer 1:
To convert the point $(0,3)$ from rectangular coordinates to polar coordinates, we need to find the values of r and θ .
Calculate r :
The formula for r is:

$$r = \sqrt{x^2 + y^2}$$
Substituting $x = 0$ and $y = 3$:

$$r = \sqrt{0^2 + 3^2} = \sqrt{9} = 3$$
Determine θ :
The angle θ can be found using the arctangent function:

$$\theta = \tan^{-1}(\frac{y}{x})$$
However, since $x = 0$ and $y = 3$, this results in an undefined value for $\tan^{-1}(0)$. Instead, we consider the position of the point in the coordinate plane. The point $(0,3)$ lies on the positive y -axis. On the positive y -axis, the angle θ is:

$$\theta = \frac{\pi}{2}$$
Thus, the polar coordinates of the point $(0,3)$ are:

$$(r, \theta) = (3, \frac{\pi}{2})$$

Answer 2:
To convert the point $(0,3)$ from rectangular coordinates to polar coordinates, we need to find the values of r and θ .
Calculate r :
The formula for r is:

$$r = \sqrt{x^2 + y^2}$$
Substituting $x = 0$ and $y = 3$:

$$r = \sqrt{0^2 + 3^2} = \sqrt{9} = 3$$
Calculate θ :
The formula for θ is:

$$\theta = \tan^{-1}(\frac{y}{x})$$
However, since $x = 0$ and $y = 3$, this point lies on the positive y -axis. Therefore, θ should be $\frac{\pi}{2}$. Thus, the polar coordinates of the point $(0,3)$ are:

$$(r, \theta) = (3, \frac{\pi}{2})$$

Answer 3:
To convert the point $(0,3)$ from rectangular coordinates to polar coordinates, we need to find r and θ .
Calculate r :
The formula for r is:

$$r = \sqrt{x^2 + y^2}$$
Substituting $x = 0$ and $y = 3$:

$$r = \sqrt{0^2 + 3^2} = \sqrt{9} = 3$$
Calculate θ :
The formula for θ is:

$$\theta = \tan^{-1}(\frac{y}{x})$$
However, since $x = 0$ and $y = 3$, this results in an undefined value for $\tan^{-1}(0)$. Instead, we consider the position of the point in the coordinate plane. The point $(0,3)$ lies on the positive y -axis. On the y -axis, the angle θ is 90° (or $\frac{\pi}{2}$ radians). Therefore, the polar coordinates of the point $(0,3)$ are:

$$(r, \theta) = (3, \frac{\pi}{2})$$

Answer 4:
To convert the point $(0,3)$ from rectangular coordinates to polar coordinates, we need to find the values of r and θ .
Calculate r :
The formula for r is given by:

$$r = \sqrt{x^2 + y^2}$$
Substituting $x = 0$ and $y = 3$:

$$r = \sqrt{0^2 + 3^2} = \sqrt{9} = 3$$
Determine θ :
The angle θ can be found using the arctangent function:

$$\theta = \tan^{-1}(\frac{y}{x})$$
However, since $x = 0$ and $y = 3$, this directly places the point on the positive y -axis. Therefore, the angle θ is:

$$\theta = \frac{\pi}{2}$$
Thus, the polar coordinates of the point $(0,3)$ are:

$$(r, \theta) = (3, \frac{\pi}{2})$$

Answer 5:
To convert the point $(0,3)$ from rectangular coordinates to polar coordinates, we need to find r and θ .
Calculate r :
The formula for r is:

$$r = \sqrt{x^2 + y^2}$$
Substituting $x = 0$ and $y = 3$:

$$r = \sqrt{0^2 + 3^2} = \sqrt{9} = 3$$
Calculate θ :
The formula for θ is:

$$\theta = \tan^{-1}(\frac{y}{x})$$
However, since $x = 0$, $\tan^{-1}(0)$ is undefined. This indicates that the point lies on the positive y -axis. Therefore, θ must be $\frac{\pi}{2}$. Putting it all together, the polar coordinates are:

$$(r, \theta) = (3, \frac{\pi}{2})$$

Assistant: `<answer>3, pi/2</answer>`

Figure 15: SSA No Think Example. We can see the Assistant output answer directly