

Structured Reasoning for Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) achieve strong performance by generating long chains of thought, but longer traces always introduce redundant or ineffective reasoning steps. One typical behavior is that they often perform unnecessary verification and revisions even if they have reached the correct answers. This limitation stems from the unstructured nature of reasoning trajectories and the lack of targeted supervision for critical reasoning abilities. To address this, we propose **StruCTured Reasoning (SCR)**, a framework that decouples reasoning trajectories into explicit, evaluable, and trainable components. We mainly implement SCR using a *Generate–Verify–Revise* paradigm. Specifically, we construct structured training data and apply Dynamic Termination Supervision to guide the model in deciding when to terminate reasoning. To avoid interference between learning signals for different reasoning abilities, we adopt a progressive two-stage reinforcement learning strategy: the first stage targets initial generation and self-verification, and the second stage focuses on revision. Extensive experiments on three backbone models show that SCR substantially improves reasoning efficiency and self-verification. Besides, compared with existing reasoning paradigms, it reduces output token length by up to 50%.

1 Introduction

Large language models (LLMs) achieve strong performance on complex reasoning tasks by generating long chains of thought (Min et al., 2024). Models such as OpenAI o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025), and the Qwen Thinking series (Yang et al., 2025) adopt a System-2 (Frankish, 2010) like reasoning style, producing answers through deliberate multi-step inference. This explicit reasoning paradigm improves decision quality and substantially enhances performance on mathematical reasoning (Shao et al., 2024) and code generation (Jiang et al., 2024).

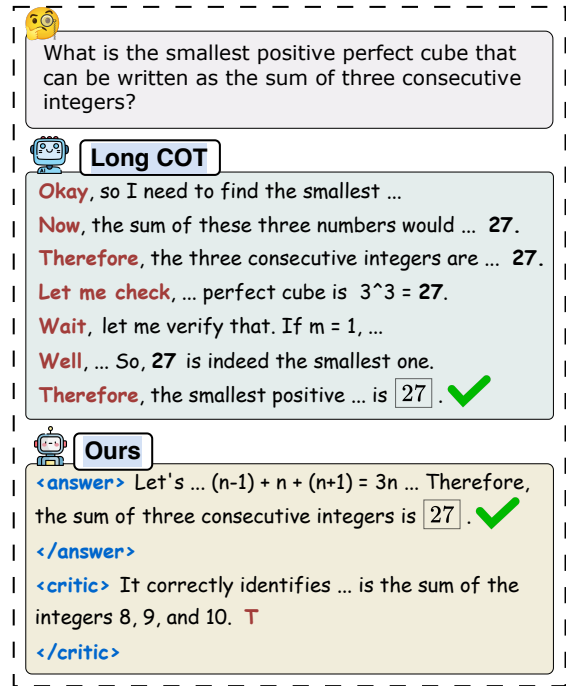


Figure 1: **Comparison of reasoning trajectories.** *Long CoT*: A single reasoning process in which all thinking operators are blended together, resulting in redundant self-verification loops even after a correct solution is reached. *Ours*: A structured reasoning trajectory with explicitly separated components, where each stage is verifiable and trainable. Generation stops once correctness is confirmed.

Despite these gains, longer reasoning traces do not necessarily guarantee high-quality reasoning (Chen et al., 2024; Luo et al., 2025), largely due to the model’s limited self-verification ability (Zhang et al., 2025). Even after producing a correct solution, it may repeatedly verify its reasoning or make unnecessary revisions (Kang et al., 2025b; Zeng et al., 2025), as illustrated in Figure 1. Most importantly, These verification and revision behaviors dominate the reasoning trajectory (Figure 2 Left) but rarely yield meaningful corrections (Figure 2 Right). When such high-frequency be-

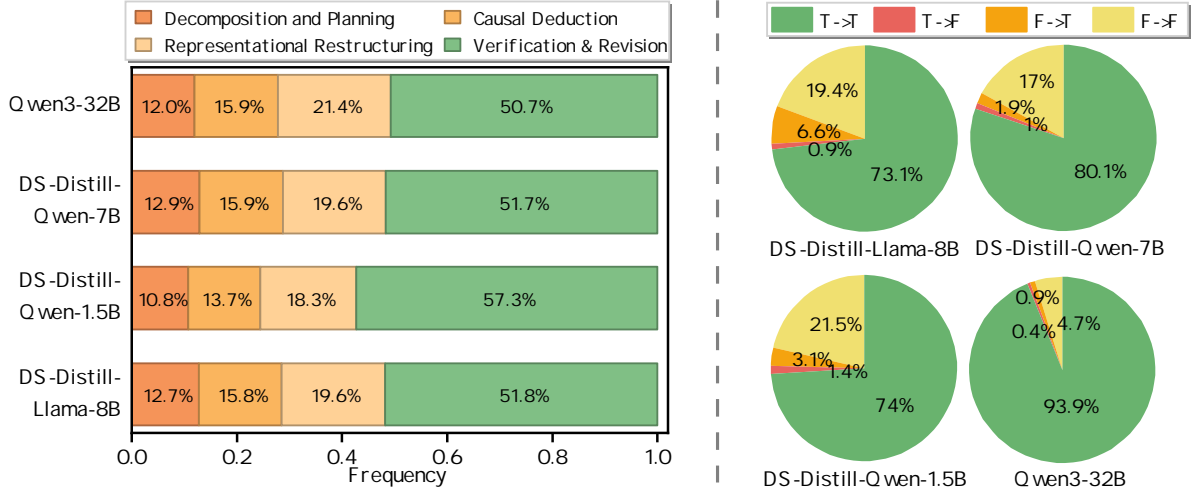


Figure 2: **Behavioral analysis of long-chain reasoning in large reasoning models.** *Left:* Frequency distribution of different thinking operators within long CoT trajectories. *Right:* Effectiveness of revision behaviors in long CoT, quantified by the transition in answer correctness pre- and post-revision.

haviors fail to produce gains, they degenerate into cognitive idling, where the model continues reasoning without substantive progress.

This issue can be traced to the unstructured nature of long reasoning trajectories under commonly used outcome-based optimization. In existing large reasoning models, different reasoning operations are intertwined within a single trajectory without clear structured boundaries (Team et al., 2025a). Furthermore, prevailing training paradigms, such as Reinforcement Learning with Verifiable Rewards (RLVR) (Kirk et al., 2024; Chu et al., 2025; Betley et al., 2025), optimize exclusively for final answer correctness. Such sparse, global reward signals provide little guidance for assessing intermediate steps, including whether verification is warranted or whether a revision effectively rectifies an error. As a result, these critical atomic abilities cannot be independently distinguished, evaluated, or trained.

Therefore, it is essential to restructure the reasoning trajectory into a structured form, where each critical reasoning ability is associated with controllable and assessable learning signals. To this end, we propose **StruCTured Reasoning (SCR)**, a framework that allows each critical ability to be independently evaluated and trained. In our implementation, we achieve this by decoupling the reasoning trajectory into the key abilities of initial generation, verification, and revision, and applying targeted optimization to each ability.

Specifically, we first construct structured reasoning training data during supervised fine-tuning

(SFT) to guide the model toward a *Generate-Verify-Revise* paradigm. We then introduce Dynamic Termination Supervision, which conditions termination decisions on the outcomes of self-verification, ensuring that verification and revision are applied appropriately. Building on this foundation, we employ a two-stage reinforcement learning strategy, optimizing each reasoning ability independently while preventing interference between their learning signals. The first stage focuses on improving initial generation and self-verification. The second stage focuses on revision and applies fine-grained learning signals to correct errors more effectively.

In summary, our contributions are three-fold:

- We reformulate long-chain reasoning as a structured reasoning trajectory, a general framework for organizing reasoning steps. In our implementation, we explicitly decouple generation, verification, and revision, enabling each critical reasoning ability to be independently evaluated and trained.
- We introduce *Dynamic Termination Supervision* in SFT, enabling the model to autonomously decide when to stop reasoning based on self-verification outcomes.
- We propose a two-stage reinforcement learning framework with fine-grained rewards to target the optimization of each reasoning ability.

2 Pilot Experiments

To examine how large reasoning models (LRM) allocate their computational budget during long-chain reasoning, we conduct a pilot study based on *Information-Processing Theory* (Simon, 1978).

This theory views problem solving as a search process driven by a set of thinking operators. Following this view, we categorize the reasoning behaviors into four types and analyze how these operators are used by LRMs. Experimental details are provided in Appendix C.

We observe that verification and revision overwhelmingly dominate the reasoning process across models. As shown in Figure 2 Left, their execution frequency substantially exceeds that of decomposition, causal deduction, and representational restructuring across four LRMs. This imbalance indicates that a large portion of the computational budget in long reasoning is allocated to post-hoc verification.

However, a closer analysis of revision outcomes reveals a clear mismatch between frequency and effectiveness. As shown in Figure 2 Right, the majority of revision steps preserve the correctness state of the initial answer. Over 90% of revisions either confirm already correct answers or leave incorrect answers unchanged ($T \rightarrow T$ or $F \rightarrow F$), while successful corrections from incorrect to correct account for only a small fraction of cases ($F \rightarrow T$). This pattern indicates that revision is rarely triggered by accurate error detection, but instead functions as repetitive confirmation.

Highlights

Overall, these results suggest that although verification and revision dominate the reasoning trajectory, they achieve a low error-correction success rate.

3 Methods

3.1 Problem Formulation

In this work, we formalize problem-solving as a single-pass generation of a structured reasoning trajectory. In this trajectory, initial generation, verification, and revision are represented as ordered components, allowing each capability to be learned, evaluated, and optimized during training.

Formally, given an input problem x , the reasoning process is represented as a trajectory:

$$y = \left(y^{(0)}, c^{(0)}, y^{(1)}, \dots, c^{(T-1)}, y^{(T)} \right), \quad (1)$$

where $y^{(0)}$ denotes the initial response, and $y^{(T)}$ is the final answer. At each iteration $t \in \{1, \dots, T\}$, the model produces a critique $c^{(t)}$, which explicitly assesses the correctness of the answer $y^{(t)}$, and when incorrect, identifies the key errors that need to be corrected. Based on this critique, the model either revises the answer to produce $y^{(t+1)}$ or keeps

the previous answer if no correction is needed. We model this process autoregressively by factorizing the probability of a reasoning trajectory as:

$$p(y | x) = p(y^{(0)} | x) \prod_{t=0}^{T-1} p(c^{(t)} | x, y^{(t)}) \cdot p(y^{(t+1)} | x, y^{(t)}, c^{(t)}). \quad (2)$$

3.2 Behavior Cloning on Structured Reasoning Trajectories

3.2.1 Data Synthesis

To construct high-quality supervision signals for structured reasoning, we develop an automated data synthesis pipeline that provides explicit training data for three critical stages.

For each problem, we first sample multiple candidate solutions to cover diverse reasoning behaviors. These candidates are automatically labeled by comparing their final answers with the ground-truth solution, yielding both correct and incorrect initial responses.

Given each labeled response, teacher models generate a corresponding critique, which either validates a correct solution or identifies errors in an incorrect one. When errors are detected, the teacher further produces a targeted revision to correct the original response. All synthesized trajectories are retained only if the final answer is verified to be correct. Further details of the data construction process are provided in the Appendix B.

Based on the synthesis pipeline described above, the SFT training data consists of two types of structured reasoning forms. The first type, **Correct-Answer Trajectories** contain cases where the initial response is correct and subsequent critiques only verify its correctness, resulting in no modification to the answer. The second type, **Correction Trajectories** contain cases where the initial response is incorrect and is corrected through one or more critique–revision rounds.

To make the internal reasoning stages explicit, each component in the training data is enclosed within paired start and end tags: initial answers are placed between `<answer>` and `</answer>`, critique content between `<critic>` and `</critic>`, and revised outputs between `<revised>` and `</revised>`. An example of this data format is shown in Box 1.

3.2.2 Supervised Fine-Tuning for Structured Reasoning

We perform Supervised Fine-Tuning (SFT) to train the model to generate structured reasoning trajectory-

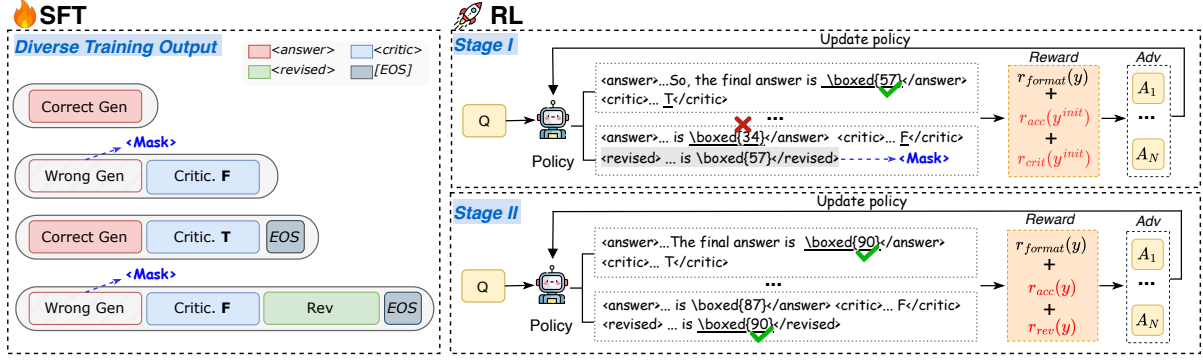


Figure 3: **Overview of the proposed training framework.** *Left:* SFT with Dynamic Termination Supervision and Selective Loss Masking. *Right:* Staged reinforcement learning with stage-specific reward designs for learning initial generation, self-verification, and revision.

ries in a single pass.

Dynamic Termination Supervision (DTS). To enable adaptive generation control, we supervise the placement of the end-of-sequence (*[EOS]*) token based on the critique outcome. For samples where the critique confirms the initial answer is correct, an *[EOS]* token is inserted immediately after the critique, signaling the model to stop generation. For samples where errors are identified, no *[EOS]* token is provided, forcing the model to continue generating a revision. This design directly supervises the model to terminate only when correctness is verified and to proceed with correction otherwise. The training data are organized in the structured form illustrated in Figure 3.

Selective Loss Masking (SLM). To prevent the model from learning incorrect reasoning patterns, we apply selective loss masking for trajectories with erroneous initial answers. Tokens in the incorrect initial answer are masked from the loss, while still serving as context for learning critique and revision.

Formally, the training objective is defined as:

$$\mathcal{L} = - \sum_{t=1}^T m_t \cdot \log P_{\theta}(y_t | x, y_{<t}), \quad (3)$$

where $m_t \in \{0, 1\}$ is a binary mask. For samples with incorrect initial answers, we set $m_t = 0$ for tokens in y^{init} and $m_t = 1$ for tokens in y^c and y^{rev} . This ensures that incorrect initial reasoning serves only as context for learning verification and correction, rather than as a supervision target.

3.3 Staged Reinforcement Learning

We apply reinforcement learning with Group Relative Policy Optimization (GRPO) to improve structured correction trajectories. The trajectory

comprises three abilities: initial generation, self-verification, and revision. Jointly optimizing all components is challenging due to entangled learning signals. Therefore, we adopt a two-stage training scheme. Stage I focuses on stabilizing initial reasoning and self-verification. Stage II then optimizes revision conditioned on reliable verification.

3.3.1 Stage I: Enhancing Initial Generation and Self-Verification.

Stage I aims to improve the model’s initial answer quality and its ability to accurately assess correctness. Following the GRPO framework, for each input x , we sample a group of outputs and update the policy using group-relative advantages.

Reward Design. Stage I adopts a rule-based reward scheme composed of three components: a format reward r_{format} , an initial accuracy reward r_{acc} , and a self-verification reward r_{crit} . These rewards supervise output structure, correctness of the initial answer, and consistency between self-verification and actual correctness, respectively.

Format Reward. It enforces a predefined output structure. Let $C_i(y) \in \{0, 1\}$ indicate whether output y satisfies constraint C_i . We define five constraints:

- C_1 : Output must include both <answer> and <critic> tag pairs;
- C_2 : <critic> section must conclude with either *T* or *F*;
- C_3 : If the critique concludes with *F*, a <revised> tag must follow;
- C_4 : Final answers in <answer> and <revised> are enclosed in `\boxed{}`;
- C_5 : All tags preserve their relative order.

The format reward is defined as:

$$r_{format}(y) = \prod_{i=1}^5 C_i(y), \quad (4)$$

which assigns a nonzero reward only to fully well-formed outputs.

Initial Accuracy Reward. This reward measures whether the model’s initial answer y^{init} matches the ground truth y^* :

$$r_{acc}(y^{init}) = \begin{cases} 1, & \text{if } y^{init} = y^*, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

Self-Verification Reward. This reward evaluates whether the model’s self-verification $crit(y) \in \{T, F\}$ correctly reflects the actual correctness of its initial answer:

$$r_{crit}(y^{init}) = \begin{cases} 1, & \text{if } (r_{acc}(y^{init}) = 1 \wedge crit(y^{init}) = T) \\ & \text{or } (r_{acc}(y^{init}) = 0 \wedge crit(y^{init}) = F), \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Total Reward. The total reward in Stage I is a weighted sum of the three components:

$$r_{total}(y) = \alpha \cdot r_{format}(y) + \beta \cdot r_{acc}(y^{init}) + \gamma \cdot r_{crit}(y^{init}). \quad (7)$$

where α , β , and γ control the relative contributions.

Optimization Objective. During Stage I, gradient updates are applied only to tokens in the `<answer>` and `<critic>` segments, while the `<revised>` block is excluded from optimization. This design isolates the learning of initial reasoning and self-verification from revision behaviors.

To implement this in GRPO, we modify the standard objective by masking the revision region. Let $y^{\backslash rev}$ denote the tokens in the `<answer>` and `<critic>` segments, excluding the `<revised>` block. Only tokens in $y^{\backslash rev}$ are included in policy optimization. The Stage I objective is then defined as:

$$J_{\text{Stage-I}}(\theta) = \mathbb{E}_{x \sim D, a_i \sim \pi_\theta(x)} \left[\frac{1}{G} \sum_{i=1}^G A_i(y_i^{\backslash rev} | x) \cdot \min \left(r_i, \text{clip}(r_i, 1 - \epsilon, 1 + \epsilon) \right) \right], \quad (8)$$

where $r_i = \frac{\pi_\theta(y_i^{\backslash rev} | x)}{\pi_{\text{old}}(y_i^{\backslash rev} | x)}$, and A_i is the group-relative advantage computed from the total reward $r_{total}(y_i)$.

3.3.2 Stage II: Optimizing Revision

After the model acquires stable initial reasoning and self-verification, Stage II optimizes the revision behavior.

Revision-Focused Rewards. Stage II adopts a compound reward that combines a format reward r_{format} , a final answer accuracy reward r_{acc} , and a revision-specific reward r_{rev} . The format reward enforces structural validity as in Stage I, while the accuracy reward evaluates the correctness of the final output.

The revision reward directly measures the effectiveness of the correction step by comparing the correctness before and after revision. It is designed according to four principles:

- (i) reward successful correction of initially incorrect answers;
- (ii) assign mild penalties to ineffective revisions
- (iii) penalize unnecessary revisions when the initial answer is already correct;
- (iv) strongly penalize harmful revisions that downgrade a correct answer.

Formally, the revision reward is defined as:

$$r_{rev}(y) = \begin{cases} \mu_1, & r_{acc}(y^{init}) = 0 \wedge r_{acc}(y) = 1 \\ \mu_2, & r_{acc}(y^{init}) = 0 \wedge r_{acc}(y) = 0 \\ \mu_3, & r_{acc}(y^{init}) = 1 \wedge r_{acc}(y) = 1 \\ \mu_4, & r_{acc}(y^{init}) = 1 \wedge r_{acc}(y) = 0 \\ 0, & \text{no revision} \end{cases} \quad (9)$$

The coefficients satisfy the ordering $\mu_1 > \mu_2 > \mu_3 > \mu_4$, reflecting the relative desirability of revision outcomes.

Total Reward. The total reward in Stage II is computed as:

$$r_{total}(y) = \nu \cdot r_{format}(y) + \eta \cdot r_{acc}(y) + \varphi \cdot r_{rev}(y). \quad (10)$$

where ν , η , and φ weight the three rewards.

Optimization Objective. In Stage II, we remove the masking strategy introduced in Stage I, enabling gradients to propagate through the entire reasoning trajectory. While retaining the same GRPO objective, the policy is now optimized over the full reasoning sequence.

4 Experiments

4.1 Experiment Setups

Models and Datasets. We conduct experiments on three widely used models: Qwen2.5-3/7B-Instruct (Team, 2024) and Llama3.1-8B-Instruct (Dubey et al., 2024). At the SFT stage, we train on the MATH (Hendrycks et al.) dataset, restricting problems to difficulty levels 3-5 and removing any overlap with the test set to prevent data

Models	Method	In-Domain					Out-of-Domain			AVG
		MATH500	AIME24	AIME25	Olympiad	AMC	ARC	MMLU	GPQA	
Qwen2.5-7B-Instruct	Base	75.60	11.00	6.00	39.70	42.05	84.39	55.80	33.84	43.55
	Self-Refine	68.40	9.33	5.67	36.44	38.67	57.51	55.90	34.85	38.35
	GRPO	77.40	11.33	10.33	42.81	50.60	85.49	57.00	37.88	46.60
	SFT + GRPO	78.80	15.33	11.67	42.81	50.84	80.63	57.80	35.86	46.72
	SCR-SFT	74.20	9.33	4.00	34.67	35.18	72.95	49.10	33.84	39.16
	SCR-Stage I	77.00	11.33	9.00	42.07	44.46	80.63	51.50	35.35	43.92
	SCR (Ours)	80.00	13.67	13.67	45.63	50.12	86.43	56.30	38.38	48.03
Qwen2.5-3B-Instruct	Base	66.60	7.00	2.33	29.93	27.95	69.20	39.60	29.80	34.05
	Self-Refine	65.60	6.67	3.00	29.78	29.52	63.57	43.50	32.32	34.25
	GRPO	70.20	6.00	5.00	33.04	35.90	81.14	44.60	36.36	39.03
	SFT + GRPO	70.00	7.33	3.67	35.11	35.06	79.01	45.30	36.36	38.98
	SCR-SFT	63.80	2.33	2.33	24.59	23.86	29.69	38.40	27.27	26.53
	SCR-Stage I	65.20	4.33	2.00	29.04	29.16	67.32	41.30	39.39	34.72
	SCR (Ours)	70.40	7.33	4.00	37.03	36.02	79.35	44.50	39.39	39.75
Llama3.1-8B-Instruct	Base	42.80	2.00	0.00	15.85	12.77	57.08	42.40	24.75	24.71
	Self-Refine	13.40	2.00	0.00	5.04	7.23	8.79	9.90	22.73	8.64
	GRPO	52.20	8.33	1.00	20.15	24.22	70.31	45.50	31.31	31.63
	SFT + GRPO	57.60	6.00	0.67	22.52	28.67	84.73	43.00	24.24	33.43
	SCR-SFT	53.40	3.00	1.33	19.11	19.04	74.06	39.10	24.75	29.22
	SCR-Stage I	55.80	4.00	1.67	23.41	26.51	82.17	42.20	31.82	33.45
	SCR (Ours)	60.80	5.33	4.67	23.70	24.46	84.56	43.20	27.27	34.25

Table 1: **SCR vs other baselines.** Compared to the Base, SCR achieves an average performance improvement of 4.48%, 5.70% and 9.54% on the three models, respectively. The best results are highlighted in bold. SCR-SFT denotes models trained with SFT only; SCR-Stage I denotes SFT followed by Stage I; SCR denotes the full two-stage RL training (default).

leakage. At the RL stage, we use DAPO-MATH¹ after filtering out all Chinese-language problems.

Evaluations. We evaluate models on both in-domain and out-of-domain benchmarks. **(i) In-domain performance.** We evaluate mathematical reasoning on AIME24, AIME25, MATH500, AMC (Li et al., 2024), and Olympiad (He et al., 2024). **(ii) Out-of-domain performance.** We assess generalization on GPQA-Diamond (GPQA) (Rein et al.), ARC (Clark et al., 2018), and a subset of MMLU-Pro (MMLU) (Wang et al., 2024b). For AIME24, AIME25, and AMC, we report $avg@10$ at temperature 0.6; for all other benchmarks, we report $pass@1$ at temperature 0.

Baselines. We compare SCR with the following baselines: (1) **Base**, the original model without fine-tuning; (2) **On-policy RL**, applying GRPO directly to the base model; (3) **Self-Refine** (Shinn et al., 2023), a prompting-based method that iteratively critiques and revises model outputs; and (4) **SFT + GRPO**, a widely adopted paradigm that first applies long-CoT supervised fine-tuning and then uses GRPO for further optimization.

Detailed experimental settings and hyperparameters are provided in Appendix A.

¹<https://huggingface.co/datasets/BytedTsinghua-SIA/DAPO-Math-17k>

4.2 Main Results

4.2.1 Comparative Analysis with Baselines

To evaluate the effectiveness of the proposed method, we compare it with representative baselines. The results are shown in Table 1.

SCR achieves the highest average performance across all three backbone models. It consistently outperforms both standard GRPO and SFT+RL approaches, demonstrating stable improvements across different parameter scales and model families.

Prompt-based self-correction is ineffective and unstable. Prompting models to reflect or revise does not yield reliable gains and can even degrade performance, indicating that self-correction cannot be robustly induced through prompt engineering alone.

SFT alone yields limited gains and can even degrade performance. For both Qwen2.5-7B and Qwen2.5-3B, the average scores fall below those of the Base models. This suggests that while SFT effectively aligns output formats, it is insufficient for improving error correction and deep reasoning, and may introduce an alignment tax (Ouyang et al., 2022). In contrast, the subsequent stages of SCR consistently recover and improve performance, in-

Backbone	Method	AVG			
		Acc.	Prec.	Rec.	F1.
Qwen2.5-7B	Base	56.31	40.92	92.96	52.09
	SCR-SFT	71.72	49.90	84.09	59.96
	SCR-Stage I	87.06	75.25	75.64	75.36
Qwen2.5-3B	Base	57.04	31.82	87.58	42.10
	SCR-SFT	65.60	36.08	80.91	45.08
	SCR-Stage I	88.23	61.71	65.02	63.20
Llama3.1-8B	Base	61.22	23.15	61.71	30.75
	SCR-SFT	63.67	29.77	77.85	40.31
	SCR-Stage I	86.66	56.88	55.43	55.81

Table 2: **Comparison of Self-Verification Performance.** The base model exhibits over-confidence and struggles to evaluate its own outputs, while SCR significantly improves self-verification accuracy.

dicating that effective reasoning alignment requires more than imitation-based supervision.

4.2.2 Analysis of Verification Capability

To evaluate the effectiveness of SCR in self-verification, we conduct experiments on three backbone models. We report Accuracy, Precision, Recall, and F1 score *averaged* over the math datasets, with results summarized in Table 2.

Base models exhibit pronounced over-confidence and lack the ability to accurately assess the correctness of their own generated answers. They achieve high Recall at the cost of low Precision and Accuracy. For example, Qwen2.5-7B Base attains a Recall of 92.96% with a large number of false positives

Applying SCR substantially enhances the model’s self-verification ability. This improvement leads to more calibrated judgments, with substantial gains in Precision and Accuracy and only a moderate reduction in Recall, resulting in a higher F1 score. The effect is most pronounced in Stage I, where self-verification is explicitly optimized.

4.3 Generalization

To evaluate whether our proposed method can generalize to other reasoning domains, we conduct experiments on commonsense reasoning benchmarks. The results are summarized in Table 1.

SCR demonstrates strong generalization capabilities. Beyond standard math tasks, our method shows substantial gains on general reasoning benchmarks. This indicates that SCR does not merely overfit to mathematical training data but fundamentally enhances the model’s robust reasoning and problem-solving abilities in complex, unseen scenarios.

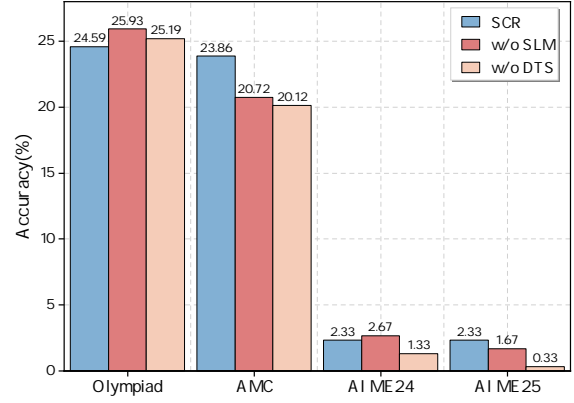


Figure 4: **Ablation of DTS and SLM during the SFT phase.** Removing DTS causes a sharp performance drop on hard benchmarks, while removing SLM leads to consistent degradation across tasks.

4.4 Analysis of Output Token Length

We evaluate output length on math benchmarks using Qwen2.5-3B-Instruct, and the results shown in Figure 6.

Our method consistently produces shorter outputs than commonly used baselines. The reduction is most pronounced on Math500, where the average output length is reduced by nearly 50%. This reduction stems from the structured reasoning process in our method. The model explicitly evaluates its initial answer and performs revision only when an error is detected; otherwise, reasoning terminates early, reducing redundant steps and overall output length.

4.5 Ablation Analysis

We conduct comprehensive ablation studies to evaluate the individual contributions of our proposed strategies.

Impact of SFT Strategies. As shown in Figure 4, we evaluate the removal of DTS and SLM on the Qwen2.5-3B-instruct backbone.

Removing DTS leads to severe performance drops on hard benchmarks. Without explicit control of the $[EOS]$ signal, the model often terminates reasoning prematurely, failing to sustain long and recursive reasoning.

Removing SLM causes consistent degradation across all benchmarks. Without masking tokens from incorrect initial answers, the model tends to reinforce erroneous reasoning patterns during training, which harms final performance.

Impact of Staged-RL. Figure 5 compares staged RL training with a unified mixed (Mix) strategy. In the Mix setting, it uses only the accuracy of the

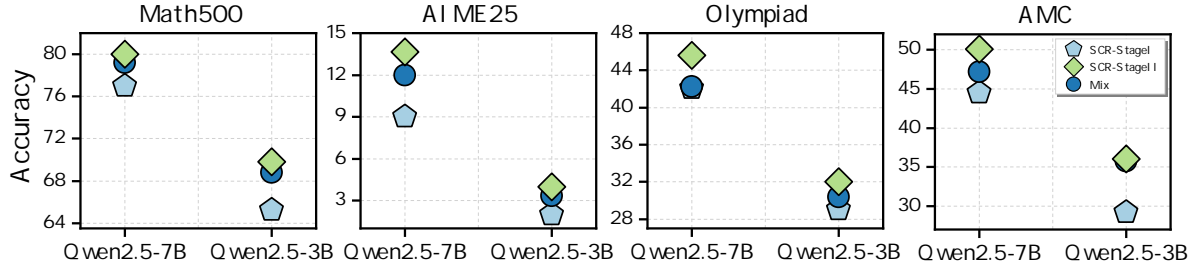


Figure 5: **Performance comparison between staged RL training and unified mixed optimization.** Progressive staged RL training leads to consistently better performance than unified mixed optimization, and the second stage further strengthens model performance.

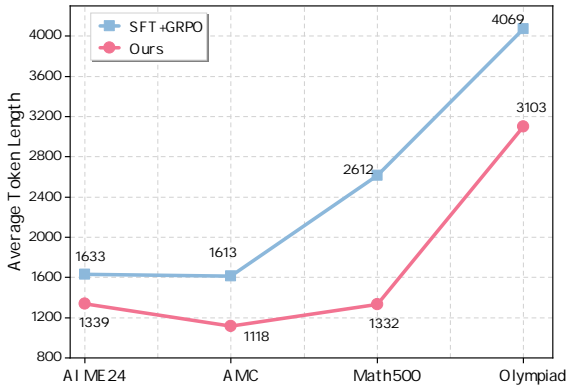


Figure 6: **Average output token length comparison.** SCR substantially reduces redundant reasoning, achieving up to 50% reduction in output length on MATH500.

final output as the reward signal.

Staged RL consistently outperforms the mixed strategy across all backbones. By optimizing different reasoning behaviors in separate stages, it reduces objective interference that arises under a single reward. Furthermore, Stage II further improves performance over Stage I on all benchmarks. For example, on AIME25, accuracy on Qwen2.5-7B increases from 9.00% to 13.67%.

5 Related Work

Enhancing Deep Thinking Capability in LLMs.

Unlike the rapid, heuristic-driven behavior of LLMs, slow-thinking reasoning systems enhance their capabilities by introducing deliberate and analytical reasoning (Snell et al., 2024). These approaches employ test-time scaling, enabling models to allocate more computation and time to reasoning before responding to challenging questions (Muennighoff et al., 2025). Reinforcement learning has recently made significant progress in enhancing the reasoning abilities of LLMs (Jaech et al., 2024; Guo et al., 2025; Team et al., 2025b). A central development is Reinforcement Learning from

Verifiable Rewards (RLVR), which replaces human feedback signals (Kirk et al., 2024; Chu et al., 2025) with automatically checkable objectives such as mathematical verification (Shao et al., 2024) and program execution (Pennino et al., 2025).

Self-correct in LLMs. Most self-correction methods in LLMs follow a two-stage paradigm, where a model generates an initial answer and then evaluates or revises it (Ganguli et al., 2023). This paradigm is widely used in complex reasoning tasks such as mathematics (Weng et al., 2023; Wang et al., 2024a) and code generation (Olausson et al., 2023), but direct prompting for self-correction is often unreliable and does not consistently improve performance (Huang et al., 2024; Tyen et al., 2024). To improve correction quality, prior work introduces external supervision, including stronger teacher models or task-specific feedback models (Xie et al., 2025; Madaan et al., 2023). However, these methods increase training cost and frequently suffer from weak generalization of correction behaviors (Du et al., 2025; Kang et al., 2025a). Recent post-training studies further show that supervised imitation alone is insufficient to induce reliable self-correction (Qu et al., 2024; Kumar et al., 2025; Kamoi et al., 2024).

6 Conclusion

In this paper, we propose StruCTured Reasoning (SCR), a framework that organizes reasoning into an explicit, assessable, and trainable form. We implement it follows the Generate–Verify–Revise paradigm. First, we apply DTS to enable the model to produce structured reasoning traces and determine reasoning termination based on self-verification. We then further enhance critical reasoning abilities through a two-stage reinforcement learning strategy. Extensive experiments on several benchmarks demonstrate that SCR consistently improves reasoning performance across tasks.

546 Limitations

547 While the structured reasoning framework naturally
548 supports multi-round verification and revision, our
549 experiments are restricted to two correction rounds.
550 This setting serves as a minimal yet sufficient con-
551 figuration to evaluate the core behaviors of critique
552 and revision. Extending to more rounds requires
553 synthesizing high-quality multi-round correction
554 trajectories, whose cost increases rapidly with the
555 number of rounds. We leave the exploration of
556 more cost-efficient strategies for multi-round data
557 construction to future work.

558 References

559 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama
560 Ahmad, Ilge Akkaya, Florencia Leoni Aleman,
561 Diogo Almeida, Janko Altenschmidt, Sam Altman,
562 Shyamal Anadkat, and 1 others. 2023. Gpt-4 techni-
563 cal report. *arXiv preprint arXiv:2303.08774*.

564 Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-
565 Betley, Xuchan Bao, Martín Soto, Nathan Labenz,
566 and Owain Evans. 2025. Emergent misalignment:
567 Narrow finetuning can produce broadly misaligned
568 llms. *arXiv preprint arXiv:2502.17424*.

569 Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He,
570 Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu,
571 Mengfei Zhou, Zhuosheng Zhang, and 1 others.
572 2024. Do not think that much for $2+3=?$ on
573 the overthinking of o1-like llms. *arXiv preprint*
574 *arXiv:2412.21187*.

575 Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang
576 Tong, Saining Xie, Dale Schuurmans, and 1 others.
577 2025. [Sft memorizes, rl generalizes: A comparative](#)
578 [study of foundation model post-training](#). *Preprint*,
579 *arXiv:2501.17161*.

580 Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot,
581 Ashish Sabharwal, Carissa Schoenick, and Oyvind
582 Tafjord. 2018. Think you have solved question an-
583 swering? try arc, the ai2 reasoning challenge. *arXiv*
584 *preprint arXiv:1803.05457*.

585 Chengyu Du, Jinyi Han, Yizhou Ying, Aili Chen,
586 Qianyu He, Haokun Zhao, Sirui Xia, Haoran Guo, Ji-
587 aqing Liang, Zulong Chen, and 1 others. 2025. Think
588 thrice before you act: Progressive thought refinement
589 in large language models. In *The Twelfth Interna-*
590 *tional Conference on Learning Representations*.

591 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey,
592 Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman,
593 Akhil Mathur, Alan Schelten, Amy Yang, Angela
594 Fan, and 1 others. 2024. The llama 3 herd of models.
595 *arXiv e-prints*, pages arXiv–2407.

596 Keith Frankish. 2010. Dual-process and dual-
597 system theories of reasoning. *Philosophy Compass*,
598 5(10):914–926.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, 599
Thomas I Liao, Kamilé Lukošiušė, Anna Chen, 600
Anna Goldie, Azalia Mirhoseini, Catherine Olsson, 601
Danny Hernandez, and 1 others. 2023. The capacity 602
for moral self-correction in large language models. 603
arXiv preprint arXiv:2302.07459. 604

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao 605
Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shi- 606
rong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. 607
Deepseek-r1: Incentivizing reasoning capability in 608
llms via reinforcement learning. *arXiv preprint* 609
arXiv:2501.12948. 610

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding 611
Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, 612
Xu Han, Yujie Huang, Yuxiang Zhang, and 1 oth- 613
ers. 2024. Olympiadbench: A challenging bench- 614
mark for promoting agi with olympiad-level bilin- 615
gual multimodal scientific problems. *arXiv preprint* 616
arXiv:2402.14008. 617

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul 618
Arora, Steven Basart, Eric Tang, Dawn Song, and 619
Jacob Steinhardt. Measuring mathematical problem 620
solving with the math dataset. In *Thirty-fifth Con-* 621
ference on Neural Information Processing Systems 622
Datasets and Benchmarks Track (Round 2). 623

Jie Huang, Xinyun Chen, Swaroop Mishra, 624
Huaixiu Steven Zheng, Adams Wei Yu, Xiny- 625
ing Song, and Denny Zhou. 2024. Large language 626
models cannot self-correct reasoning yet. In *The* 627
Twelfth International Conference on Learning 628
Representations. 629

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richard- 630
son, Ahmed El-Kishky, Aiden Low, Alec Helyar, 631
Aleksander Madry, Alex Beutel, Alex Carney, and 1 632
others. 2024. Openai o1 system card. *arXiv preprint* 633
arXiv:2412.16720. 634

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, 635
and Sunghun Kim. 2024. A survey on large lan- 636
guage models for code generation. *arXiv preprint* 637
arXiv:2406.00515. 638

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, 639
and Rui Zhang. 2024. [When can LLMs actually](#) 640
[correct their own mistakes? a critical survey of self-](#) 641
[correction of LLMs](#). *Transactions of the Association* 642
for Computational Linguistics, 12:1417–1440. 643

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, 644
and Sergey Levine. 2025a. [Unfamiliar finetuning](#) 645
[examples control how language models hallucinate](#). 646
In *Proceedings of the 2025 Conference of the Na-* 647
tions of the Americas Chapter of the Association for 648
Computational Linguistics: Human Language Tech- 649
nologies (Volume 1: Long Papers), pages 3600–3612, 650
Albuquerque, New Mexico. Association for Compu- 651
tational Linguistics. 652

Liwei Kang, Yue Deng, Yao Xiao, Zhanfeng Mo, 653
Wee Sun Lee, and Lidong Bing. 2025b. First try 654
matters: Revisiting the role of reflection in reasoning 655
models. *arXiv preprint arXiv:2510.08308*. 656

657	Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity . In <i>The Twelfth International Conference on Learning Representations</i> .	reasoning to code: Grpo optimization for underrepresented languages. <i>arXiv preprint arXiv:2506.11027</i> .	714 715
663	Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, and 1 others. 2025. Training language models to self-correct via reinforcement learning. In <i>The Twelfth International Conference on Learning Representations</i> .	Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. 2024. Recursive introspection: Teaching language model agents how to self-improve . In <i>The Thirty-eighth Annual Conference on Neural Information Processing Systems</i> .	716 717 718 719 720
669	Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, and 1 others. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. <i>Hugging Face repository</i> , 13(9):9.	David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In <i>First Conference on Language Modeling</i> .	721 722 723 724 725
676	Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. <i>arXiv preprint arXiv:2501.12570</i> .	Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models . <i>Preprint</i> , arXiv:2402.03300.	726 727 728 729 730 731
681	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In <i>Proceedings of the 37th International Conference on Neural Information Processing Systems</i> , NIPS '23, Red Hook, NY, USA. Curran Associates Inc.	Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In <i>Proceedings of the Twentieth European Conference on Computer Systems</i> , pages 1279–1297.	732 733 734 735 736 737
691	Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, and 1 others. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. <i>arXiv preprint arXiv:2412.09413</i> .	Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. <i>Advances in Neural Information Processing Systems</i> , 36:8634–8652.	738 739 740 741 742
697	Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling . <i>Preprint</i> , arXiv:2501.19393.	Herbert A Simon. 1978. <i>Information-processing theory of human problem solving</i> . Erlbaum Hillsdale, NJ.	743 744
702	Theo X. Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. 2023. Demystifying gpt self-repair for code generation . <i>CoRR</i> , abs/2306.09896.	Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters . <i>Preprint</i> , arXiv:2408.03314.	745 746 747 748
706	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. <i>Advances in neural information processing systems</i> , 35:27730–27744.	Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025a. Kimi k1. 5: Scaling reinforcement learning with llms. <i>arXiv preprint arXiv:2501.12599</i> .	749 750 751 752 753
712	Federico Pennino, Bianca Raimondi, Massimo Rondelli, Andrea Gurioli, and Maurizio Gabbriellini. 2025. From	Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, and 1 others. 2025b. Kimi k1.5: Scaling reinforcement learning with llms . <i>Preprint</i> , arXiv:2501.12599.	754 755 756 757
		Qwen Team. 2024. Qwen2.5: A party of foundation models .	758 759
		Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location . In <i>Findings of the Association for Computational Linguistics: ACL 2024</i> , pages 13894–13908, Bangkok, Thailand. Association for Computational Linguistics.	760 761 762 763 764 765 766

- 767 Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai
768 Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui.
769 2024a. Math-shepherd: Verify and reinforce LLMs
770 step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand.
- 774 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni,
775 Abhranil Chandra, Shiguang Guo, Weiming Ren,
776 Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others.
777 2024b. Mmlu-pro: A more robust and challenging
778 multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.
- 781 Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He,
782 Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao.
783 2023. [Large language models are better reasoners with self-verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore. Association for Computational Linguistics.
- 788 Zhihui Xie, Jie chen, Liyu Chen, Weichao Mao, Jingjing
789 Xu, and Lingpeng Kong. 2025. [Teaching language models to critique via reinforcement learning](#). In *ICLR 2025 Third Workshop on Deep Learning for Code*.
- 793 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang,
794 Binyuan Hui, Bo Zheng, Bowen Yu, Chang
795 Gao, Chengen Huang, Chenxu Lv, and 1 others.
796 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- 798 Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yun-
799 hua Zhou, and Xipeng Qiu. 2025. Revisiting the
800 test-time scaling of o1-like models: Do they truly
801 possess test-time scaling capabilities? *arXiv preprint arXiv:2502.12215*.
- 803 Qingjie Zhang, Di Wang, Haoting Qian, Yiming Li,
804 Tianwei Zhang, Minlie Huang, Ke Xu, Hewu Li, Liu
805 Yan, and Han Qiu. 2025. Understanding the dark
806 side of llms’ intrinsic self-correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27066–27101.
- 810 Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan
811 Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma.
812 2024. [Llamafactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

A Experiments Details

Hardware and Software Platform. All supervised fine-tuning (SFT) experiments are conducted on servers equipped with eight NVIDIA A800-SXM4 GPUs, each with 40GB memory, using the LLaMA-Factory (Zheng et al., 2024) framework. Reinforcement learning (RL) experiments are performed on servers with eight NVIDIA H20 GPUs, using the Verl (Sheng et al., 2025) framework.

SFT Training Configuration. Supervised fine-tuning is performed with full-parameter optimization for one epoch. The maximum sequence length is set to 14,000 tokens. Training uses a per-device batch size of 1 with gradient accumulation over 8 steps. The model is optimized with Adam using a learning rate of 1×10^{-6} , together with a cosine learning rate schedule and a warmup ratio of 0.1. We enable bf16 precision and employ DeepSpeed ZeRO-3 for efficient training.

RL Training Configuration. We train the model using the GRPO algorithm for 200 optimization steps. During training, rollouts are generated using temperature sampling with $\tau = 1.0$, and end-of-sequence tokens are enforced. We adopt a grouped sampling strategy with a group size of $G = 8$ responses per problem.

For efficient rollout generation, we use vLLM with paged attention, enabling batched decoding with 128 rollout slots. The maximum prompt length is set to 1,024 tokens, and the maximum response length is 8,000 tokens.

The actor model is optimized using the Adam optimizer with a learning rate of 1×10^{-6} and a batch size of 64. For policy optimization, the clipping range is set to $[0.01, 0.99]$.

We employ different reward designs. In Stage I, reward weights are set to $\alpha = 0.1$, $\beta = 1.0$, and $\gamma = 1.0$, balancing formatting quality, initial answer accuracy, and critic accuracy. In Stage II, reward weights are set to $\eta = 1.0$ and $\varphi = 1.0$. To encourage successful revisions, we further set $\mu_1 = -0.1$, $\mu_2 = -0.3$, $\mu_3 = -0.5$, and $\mu_4 = -0.5$.

Inference Configuration. During inference, we use vLLM for efficient decoding. The sampling temperature is set to 0 with top- p sampling at 0.95. The maximum generation length is 24,000 tokens.

For AIME24, AIME25 and AMC evaluations, the sampling temperature is set to 1.0.

Training Data of SFT + GRPO. The *SFT + GRPO* baseline is trained on long chain-of-thought trajectories distilled from Qwen3-32B, where the model produces correct solutions on MATH levels 3-5. To control for data scale, we retain 7K training samples for SFT across all methods.

All methods use the same dataset during the RL stage to ensure a fair comparison. All RL experiments are performed on the DAPO-Math dataset.

B Structured Reasoning Data construction of SCR.

We construct structured supervision data through an automated synthesis pipeline that explicitly decomposes reasoning into three stages: initial response, critique, and revision. All prompts used in this section are listed in Box 4 and 5.

Initial Response Sampling and Categorization. For each problem, we generate multiple candidate solutions to capture diverse reasoning patterns. To achieve this, we employ multiple models to sample responses, ensuring a variety of reasoning strategies. The candidates are automatically categorized as correct or incorrect based on the ground-truth answers. These responses serve as the foundation for subsequent critique and revision construction, rather than being used directly for training.

Critique Construction. To supervise self-evaluation, each problem is paired with both correct and incorrect candidate responses. Teacher models are then used to generate critiques for these answer pairs. Only critiques whose evaluative conclusions align with the ground-truth correctness are retained, ensuring reliable feedback for the model to learn how to identify reasoning errors.

Revision Trajectory Construction. Revision trajectories are generated using a hierarchical strategy to encourage the model to correct initial errors with minimal modifications. First, *Refinement-based Revisions* attempt localized edits to fix mistakes while preserving valid reasoning in the initial answer. If refinement fails to produce a correct answer, *Replacement-based Revisions* generate a new solution along an alternative reasoning path. All trajectories are verified to ensure the final answer is correct.

Based on the synthesis pipeline described above, each component is enclosed within paired start and

Method	Math500				AIME24				Olympiad				AMC			
	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
Qwen2.5-7B-Instruct																
Base	76.95	79.91	92.84	85.89	52.33	18.75	100.00	31.58	56.15	47.41	95.52	63.37	59.64	51.10	93.12	65.99
SCR-SFT	87.35	87.28	99.43	92.96	69.53	29.17	73.68	41.79	62.29	57.72	99.07	72.95	69.51	61.35	78.28	68.79
SCR-Stage I	87.02	90.82	93.13	91.96	92.86	65.71	71.88	68.66	84.14	81.31	83.04	82.17	82.40	80.81	80.16	80.48
Qwen2.5-3B-Instruct																
Base	71.80	71.92	94.59	81.71	52.00	9.80	71.43	17.24	50.37	36.51	89.11	51.80	53.37	35.62	82.76	49.81
SCR-SFT	78.57	78.16	100.00	87.74	65.00	5.00	33.33	8.70	58.70	55.12	100.00	71.07	57.08	37.60	71.21	49.21
SCR-Stage I	80.63	82.55	90.85	86.50	96.61	40.00	50.00	44.44	82.76	73.63	75.51	74.56	83.15	72.37	68.75	70.51
Llama3.1-8B-Instruct																
Base	68.75	59.86	82.63	69.43	55.67	2.96	66.67	5.67	71.13	32.95	80.19	46.70	57.19	20.00	79.05	31.92
SCR-SFT	69.83	68.32	98.86	80.81	63.79	7.14	50.00	12.50	48.03	38.41	96.80	55.00	59.66	28.71	76.92	41.81
SCR-Stage I	78.36	79.93	85.66	82.70	94.33	50.00	35.29	41.38	81.87	69.01	62.42	65.55	81.39	85.43	93.75	89.40

Table 3: **The detailed performance of self-evaluation capability on math tasks.** SCR maintains strong self-evaluation ability even on challenging problems.

Stage	Settings	Math500	AIME24	AIME25	Olympiad	AMC	AVG
SCR-Stage I	$\beta = 1.0, \gamma = 1.0$	65.20	4.33	2.00	29.04	29.16	25.95
	$\beta = 0.5, \gamma = 1.0$	40.20	1.33	0.33	13.48	10.24	13.12
	$\beta = 1.0, \gamma = 0.5$	65.00	6.33	4.00	30.81	32.53	27.73
SCR-Stage II	$\eta = 1.0, \varphi = 1.0$	70.40	7.33	4.00	37.03	36.02	30.96
	$\eta = 1.0, \varphi = 2.0$	68.20	7.67	4.00	32.30	37.11	29.86
	$\eta = 2.0, \varphi = 1.0$	68.40	7.00	5.33	32.89	34.94	29.71

Table 4: **Impact of reward weight coefficients on performance during Stage I and Stage II training.**

end tags: initial answers are placed between `<answer>` and `</answer>`, critique content between `<critic>` and `</critic>`, and revised outputs between `<revised>` and `</revised>`.

C Pilot Experiment Details

In Section 2, we analyze the long CoT reasoning trajectories produced by four representative deep reasoning models: DeepSeek-R1-Distill-Qwen-7B², DeepSeek-R1-Distill-Qwen-1.5B³, DeepSeek-R1-Distill-Llama-8B⁴, and Qwen3-32B⁵. All models are evaluated on the MATH500 dataset under a deterministic decoding setting with temperature set to zero.

To characterize the internal reasoning dynamics, we first define four categories of thinking operators that capture the major functional roles observed in long CoT reasoning. Using GPT-4.1-2025-04-14 (Achiam et al., 2023) as a teacher annotator,

²<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B>

³<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B>

⁴<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B>

⁵<https://huggingface.co/Qwen/Qwen3-32B>

we label each reasoning trajectory at the operator level and count the occurrence frequency of each operator. The detailed definitions of the operators and the annotation prompts are provided in the Box 6 and 7.

Beyond operator frequency analysis, we further investigate the effectiveness of the *Verification & Revision* operator, which plays a central role in long-form reasoning. For each reasoning trajectory, we extract the answer path that leads to the final prediction and compare the model’s answer before and after revision against the ground truth. Based on this comparison, each instance is categorized into one of four outcomes: correct-to-correct (T→T), correct-to-incorrect (T→F), incorrect-to-incorrect (F→F), and incorrect-to-correct (F→T). This analysis allows us to quantify whether verification and revision behaviors meaningfully improve answer correctness or merely increase reasoning length without yielding effective corrections.

D Further Analysis

D.1 Self-evaluation Capability Analysis

In Section 4, we report the average self-evaluation performance of our method on math tasks. To fur-

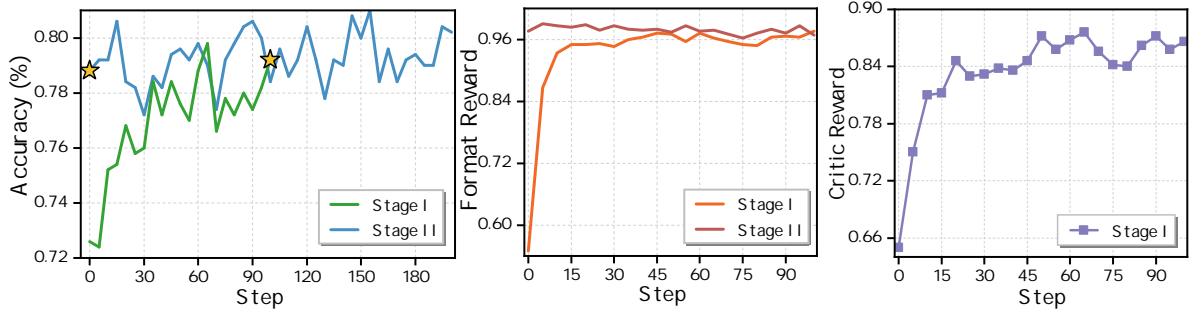


Figure 7: **Training dynamics of Qwen2.5-7B-Instruct.** From left to right, we report validation accuracy, format score, and self-evaluation reward over training.

958 ther examine how self-evaluation behavior varies
 959 with task difficulty, we analyze SCR under differ-
 960 ent difficulty settings, with the results summarized
 961 in Table 3.

962 **SCR-Stage I substantially strengthens the**
 963 **model’s self-evaluation ability across task dif-**
 964 **ficulties.** This improvement is mainly reflected in
 965 large gains in precision, indicating that the model
 966 learns to judge when it should answer. On hard
 967 benchmarks such as Olympiad and AIME24, where
 968 base models suffer severe precision drops, SCR-
 969 Stage I consistently restores high precision, show-
 970 ing effective suppression of hallucinated or low-
 971 confidence reasoning.

972 **The gain in precision does not come at the cost**
 973 **of recall.** Unlike conservative strategies that im-
 974 prove precision by avoiding answers, SCR-Stage I
 975 maintains competitive recall, resulting in balanced
 976 improvements in F1 score. This suggests that the
 977 model acquires calibrated self-evaluation rather
 978 than simply becoming hesitant.

979 **The improvement in self-evaluation general-**
 980 **izes across models and difficulty levels.** Consis-
 981 tent precision gains are observed on both strong and
 982 weak models, from Math500 to competition-level
 983 benchmarks, indicating that SCR-Stage I teaches a
 984 reliable “know-when-to-answer” behavior indepen-
 985 dent of model scale.

986 **SCR-Stage I also enables small models to**
 987 **outperform larger base models.** With struc-
 988 tured self-evaluation, Qwen2.5-3B-Instruct sur-
 989 passes Llama3.1-8B-Instruct (Base) on Math500
 990 and AMC, despite having less than half the param-
 991 eters, demonstrating strong efficiency and scalability
 992 benefits.

993 D.2 Important Parameters Analysis

994 Following Eq 7 and Eq 10, different reward weights
 995 are used in the RL stages. Since the model already

996 learns to follow the output format during SFT, as
 997 also reflected by the high format scores during the
 998 RL stage (Figure 7 Middle), we analyze only the
 999 remaining reward-weight parameters on Qwen2.5-
 1000 3B-Instruct backbone. Results are reported in Ta-
 1001 ble 4.

1002 In the stage I training, **the accuracy signal plays**
 1003 **a dominant role.** When the weight of the accu-
 1004 racy reward is reduced to $\beta = 0.5$, performance
 1005 drops sharply, with the average score falling to
 1006 13.12%. This indicates that learning to produce
 1007 correct answers remains the primary driver of ef-
 1008 fective training and should not be weakened.

1009 Interestingly, the best performance is achieved
 1010 when the self-evaluation reward weight is reduced
 1011 to $\gamma = 0.5$, outperforming the setting with equal
 1012 weights. This suggests that while self-evaluation
 1013 is essential, it functions best as an auxiliary regu-
 1014 larizer. Assigning it an overly large weight may
 1015 interfere with the main objective and distract the
 1016 model from accurate reasoning.

1017 In the stage II training, we explore the interac-
 1018 tion between the accuracy reward and the revision
 1019 reward. Unlike Stage I, optimal performance is
 1020 achieved under a more balanced configuration. The
 1021 highest average accuracy is obtained when $\eta = 1.0$
 1022 and $\varphi = 1.0$.

1023 Deviating from this balance leads to perfor-
 1024 mance degradation. Over-emphasizing either the
 1025 revision reward or the final accuracy reward re-
 1026 sults in a mild decline. This indicates that effective
 1027 self-revision requires a careful trade-off between
 1028 encouraging corrective behavior and preserving fi-
 1029 nal answer correctness, and that over-optimizing
 1030 either objective can be detrimental.

1031 D.3 Training Dynamics

1032 We report the monitored metrics during RL train-
 1033 ing. The results show that Stage II further improves

1034 accuracy on top of Stage I, while Stage I substan-
1035 tially enhances the model’s self-evaluation capa-
1036 bility. The detailed training metrics as shown in
1037 [7](#).

1038 **E Prompt Templates.**

1039 We provide the prompt templates used in our ex-
1040 periments, including:

- 1041 • The prompt used in SCR
- 1042 • The prompt that guide the teacher model to gen-
1043 erate structured critique signals
- 1044 • The prompt that guide the teacher model to gen-
1045 erate refinement answer
- 1046 • The prompt used for analyzing reasoning tra-
1047 jectories by measuring the frequency of different
1048 cognitive operators
- 1049 • The prompt designed to extract the answers of
1050 deep reasoning models

Training Data Format of SCR: Correction Trajectories

/Input**/**

1. People numbered 1, 2, \dots , 2015 are arranged in a line, and a position-swapping game is played among them, with the rule that each swap can only occur between adjacent individuals. Now, the person numbered 100 and the person numbered 1000 are to swap positions, with the minimum number of swaps required being $\quad\quad\quad$ times.

/Output**/**

<answer>

Let's reason through this problem:

We need to find the minimum number of swaps needed for person 100 and person 1000 to swap their positions.

The key here is understanding that every swap moves two people forward or backward by one place. Specifically, we want to move person 100 and person 1000 relative to each other by 900 positions (from 100 to 1000). Since each swap affects two positions at once,

(... more reasoning tokens omitted for brevity)

Considering efficiency and the simplicity of the problem, we calculate:

$$\text{Minimum swaps required} = 900 + (\text{Efficiency savings})$$

Given our targeted statement on "minimum", we assume efficiency yields ideal movement, thus aligning our answer closely with Calculation 1:

900

</answer>

</critic>

The key mistake is that the answer incorrectly equates the number of swaps with the difference in the initial positions of the two individuals.

This is wrong because it only accounts for the movement of one person to the other's position, ignoring the movement required for the second person. To swap person A (at position 100) and person B (at position 1000), *(... more reasoning tokens omitted for brevity)*

F.

</critic>

<revised>

To improve the answer, it's important to recognize the full process for swapping two individuals who are not adjacent.

(... more reasoning tokens omitted for brevity)

This is a distance of 899 positions, requiring 899 swaps. Each swap moves person 1000 one spot towards position 100.

Thus, the correct minimum number of adjacent swaps required is the sum:

$$900 \text{ (to move 100 to 1000)} + 899 \text{ (to move 1000 to 100)} = 1799$$

1799

<revised>

Training Data Format of SCR: Correct Answer Trajectories

/Input**/**

The lengths in feet of three pieces of timber are 48, 72, and 40. The sawmill operator needs to cut the timber into logs of equal length with no waste. How many feet long is the greatest possible length she can cut?

/Output**/**

<answer>

To find the greatest possible length that can be used to cut the timber into equal pieces with no waste, we need to find the greatest common divisor (GCD) of the lengths of the three pieces of timber. The GCD is the largest number that divides all three lengths without leaving a remainder.

The lengths of the timber are 48, 72, and 40 feet.

First, we find the prime factorization of each length:

- $48 = 2^4 \cdot 3$
- $72 = 2^3 \cdot 3^2$
- $40 = 2^3 \cdot 5$

Next, we identify the common factors among the prime factorizations. The common factors are 2^3 (the smallest power of 2 present in all three factorizations) and 1 (since there are no other common factors).

Therefore, the GCD of 48, 72, and 40 is

$$2^3 \cdot 1 = 8$$

So, the greatest possible length she can cut is 8 feet.

8

</answer>

<critic>

The answer is correct.

The model correctly identifies that the problem requires finding the greatest common divisor (GCD) of the three lengths. The calculation of the GCD using prime factorization is accurate. **T**

</critic>

Prompt Used in SCR

You are a helpful AI assistant.

For each question, you must first solve the problem and put your complete solution inside `<answer>...</answer>`. The final result must be wrapped in `\boxed{}`.

Then you must evaluate your own solution inside `<critic>...</critic>`. In this block, you must first give your evaluation and reasoning. At the end of the `<critic>`, you must give the final judgment using only one single symbol: T or F. T means the answer is correct. F means the answer is incorrect.

If the final judgment is F, you must give a corrected solution inside `<revised>...</revised>`, and the final result must also be wrapped in `\boxed{}`.

If the final judgment is T, you must stop and give no further output.

1053

Teacher Prompt for Generating Critique Information

/* System Prompt */

You are a careful and strict evaluator.

Given a question and a model-generated answer, judge whether the answer is correct.

Evaluation rules:

- Focus on correctness, not writing style.
- If the answer is correct, briefly explain why.
- If the answer is incorrect:
 - Clearly state the key mistake.
 - Explain why it is wrong.
 - Give a short, actionable suggestion to fix it.
- Be concise and precise.
- Do NOT restate the full answer or redo the full solution.

End your response with a single letter on a new line:

T if the answer is correct.

F if the answer is incorrect.

/* User Prompt */

Question: {here is the question}

Answer: {here is the answer}

1054

Teacher Prompt for Generating Correction Answer

/* System Prompt */

You are a helpful AI assistant.

You will be given a question, an initial answer, and evaluation feedback on that answer.

Your task is to use the evaluation feedback to modify the initial answer so that the result becomes correct.

Describe the modification process in a natural and explanatory way.

Explain what needs to be corrected and how the correction is carried out, as part of the reasoning.

Do not use explicit section titles or meta descriptions such as “refinement process” or “final answer”.

Do not refer to the evaluation itself.

After completing the modification and reasoning, present the corrected final result.

Wrap the final answer in `\boxed{ . . . }`.

/* User Prompt */

Question: {here is the question}

Answer: {here is the answer}

Evaluation: {here is the evaluation feedback}

Prompt for Measuring the Frequency of Different Thinking Operators

You are a **Cognitive Operator Analyzer & Counter**. The user will provide a QA pair. Your task has **TWO PHASES**:

Phase 1: Detailed Analysis - Analyze the Answer (A) in detail, identifying occurrences of different cognitive operators. - You may explain and reason freely. - Identify where and how each operator appears.

Phase 2: Final Operator Count (Strict Output) - After analysis, output a clean operator count summary in strict format for automated parsing.

Operators to Count:

- 1) **DECOMP_PLAN (Decomposition & Planning Operator)**: One coherent planning episode; sub-steps do not count separately. Multiple independent plans count separately.
- 2) **CAUSAL_INFER (Causal Inference Operator)**: One per causal chain; multiple connectors in the same chain count as one; new independent chains count separately.
- 3) **MONITOR (Monitoring Operator)**: Each evaluation or self-check counts as one instance.
- 4) **BACKTRACK (Backtracking / Step-Revision Operator)**: Each revision or correction of a reasoning step counts once.
- 5) **REPR_REFRAME (Representation Reframing Operator)**: Each distinct reframing or perspective shift counts once.

Counting Rules: - Analyze only the Answer (A), ignore the Question (Q). - Count each operator action once. - Multiple operators can appear in the same sentence if distinct. - Assign each instance to the most dominant operator. - Do not include explanations in the final output.

Final Output Format (Mandatory): Your output must contain only the following five lines:

```
COUNT_DECOMP_PLAN:<int>
COUNT_CAUSAL_INFER:<int>
COUNT_MONITOR:<int>
COUNT_BACKTRACK:<int>
COUNT_REPR_REFRAME:<int>
```

Prompt for Extracting the Final Answer

You are an “**Answer-Level Trajectory Recorder**”.

The user will provide a QA pair. Your goal is to analyze **ONLY the Answer (A)** and extract the full trajectory of **ANSWER-LEVEL STATES** for the **TARGET** quantity asked by the question. This task analyzes how the model’s ANSWER changes over time.

Phase 1: Analysis (Free-Form) - First, read the Question (Q) **ONLY** to identify the **TARGET** quantity (what the question ultimately asks the model to answer). - Then analyze the Answer (A) and reason about how the model treats different values or expressions as the ANSWER to that TARGET over time. - In this phase, distinguish answer-level commitments from intermediate computations, and identify moments where the model:

- proposes a candidate final answer
- temporarily commits to an answer
- revises or replaces a previously stated answer
- re-confirms or restates the same answer

Phase 2: Final Output (Strict) - After the analysis, output the **ANSWER-LEVEL TRAJECTORY** in chronological order.

What Counts as an Answer-Level State An answer-level state **MUST** satisfy **BOTH** conditions:

- (1) **Target Alignment:** Refers to the TARGET quantity asked by Q (e.g., “the value of $f(-2)+f(-1)+f(0)$ ”, “the final common fraction”, “the correct option”).
- (2) **Answer Commitment:** The model presents it as an answer, conclusion, or result for the TARGET, including:
 - “the answer is ...”
 - “the value is ...”
 - “thus ... equals ...”
 - “final answer: ...”
 - an explicit restatement or confirmation of a previously given answer

Important: A statement counts even if it later turns out to be wrong.

What Must Be Filtered Out Do **NOT** extract:

- Sub-results used only to compute the TARGET (e.g., $f(-2)=2$, $f(-1)=5/3$, $f(0)=1$)
- Pure intermediate calculations or simplification steps
- Values not presented as the answer to the TARGET

Even if boxed or emphasized, they are NOT answer-level states unless treated as the answer.

Trajectory Rules

- Extract answer-level states strictly in the order they appear in A.
- Do NOT merge, deduplicate, or normalize answers.
- If the same answer is stated or confirmed multiple times, record EACH occurrence separately.
- If multiple different candidate final answers appear, record ALL of them.
- Do NOT invent answer states that do not explicitly appear in A.

Final Output Format (Mandatory) Your output must contain ONLY **boxed answer-level states**, one per line. No explanations, headers, numbering, or blank lines.

`\boxed{}`