RLKD: Distilling LLMs' Reasoning via Reinforcement Learning

Shicheng Xu[1,2], Liang Pang[1*], Yunchang Zhu[3], Jia Gu[1,2], Zihao Wei[1,2], Jingcheng Deng[1,2], Feiyang Pan[3], Huawei Shen[1], Xueqi Cheng[1]

[1]State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Huawei Inc.
{xushicheng21s,pangliang,shenhuawei,cxq}@ict.ac.cn, zhuyunchang@huawei.com, pfy824@gmail.com,

Abstract

Distilling reasoning paths from teacher to student models via supervised fine-tuning (SFT) provides a shortcut for improving the reasoning ability of the smaller Large Language Models (LLMs). However, the reasoning paths generated by teacher models often reflect only surface-level traces of their underlying authentic reasoning. Insights from cognitive neuroscience suggest that authentic reasoning involves a complex interweaving between meta-reasoning that selects the appropriate sub-problem from multiple candidates, and solving, which addresses the sub-problem. It means that authentic reasoning has implicit multi-branch structure. Supervised fine-tuning collapses this rich structure into a flat sequence of token prediction in teacher's reasoning path, which cannot distill this structure to student. To address this limitation, we propose RLKD, a reinforcement learning (RL)-based distillation framework guided by a novel Generative Structure Reward Model (GSRM). Our GSRM converts the reasoning path into multiple meta-reasoning-solving steps and gives the reward to measure the alignment between the reasoning structures of student and teacher. Our RLKD combines this reward with RL, enables the student LLM to internalize the teacher's implicit multi-branch structure in authentic reasoning, rather than merely mimicking fixed teacher's output paths. Experiments show that RLKD, even when trained on only 0.1% of the data under an RL-only regime, surpasses the performance of standard SFT-RL pipelines and further unleashes the potential reasoning ability of the student LLM than SFT-based distillation.

Code — https://github.com/xsc1234/RLKD

1 Introduction

Recently, Large Language Models (LLMs) have demonstrated impressive abilities on complex reasoning tasks (Pan et al. 2025; Jiaqi et al. 2025; Xu et al. 2025b) via generating long reasoning paths (Havrilla et al. 2024; Wei et al. 2022) such as Deepseek-R1 (Guo et al. 2025). However, high training costs (Guo et al. 2025; Xu et al. 2025a) and the strong base model (Chu et al. 2025; Yue et al. 2025) are required in order for LLMs to emerge with this excellent capability, which prevents this reasoning capability from being explored by resource-constrained teams in developing their LLMs. To solve this challenge, supervised fine-tuning (SFT) on the reasoning paths generated from LLMs with powerful reasoning capabilities provides a shortcut and efficient method to make the smaller LLMs generate the long reasoning paths and achieve significant improvement (Guo et al. 2025; Zhang et al. 2025; Face 2025; Wen et al. 2025). Despite this advance, some studies find SFT-distilled reasoning LLMs are trapped in rigid imitation rather than authentic reasoning. Purely mimicking the teacher's reasoning paths can leave the student LLM "unthinking": it replicates the surface form of the reasoning steps yet still makes errors on key underlying steps (Chen et al. 2025; Dai et al. 2024).

To analyze and solve this phenomenon, we introduce the concepts from human cognitive neuroscience to rethink the definition of authentic reasoning of LLMs, which consists of two parts: one is meta-reasoning and the other is solving (Cox and Raja 2007; Russell and Wefald 1991). Specifically, answering a complex problem involves multiple steps and each step consists of a meta-reasoning phase that determines the specific sub-problem from multiple potential sub-problems – followed by a solving phase that executes or answers the specific determined sub-problem. From this perspective, although the generated reasoning content is a definitive path, each step on this path is actually determined by meta-reasoning from multiple candidate states. Therefore, the complex interweaving between meta-reasoning and solving constitutes the authentic reasoning, in which both generated reasoning path and multiple other potential paths form the implicit multi-branch structure (Figure 1 (a)). A critical challenge in SFT-based distillation is that it trains a student LLM to imitate the teacher's output sequence token-by-token with cross-entropy loss. Therefore, SFT collapses this rich implicit multi-branch structure in teacher into a flat sequence of token prediction to memorize only the teacher's generated path (Figure 1 (b)) while fails to learn how to sample the path from other potential paths.

To solve this problem, it is important to provide the learning signal in distillation that clearly organizes multi-step meta-reasoning and solving in the reasoning paths, which is the core of implicit multi-branch structure. This is a highly semantic supervision that cannot be accomplished by token-level SFT and requires reinforcement learning (RL) (Ouyang et al. 2022; Havrilla et al. 2024), so we propose RLKD, the first reinforcement learning-based knowledge distillation method for LLM's reasoning. In RLKD, we design Genera-

*Corresponding Author

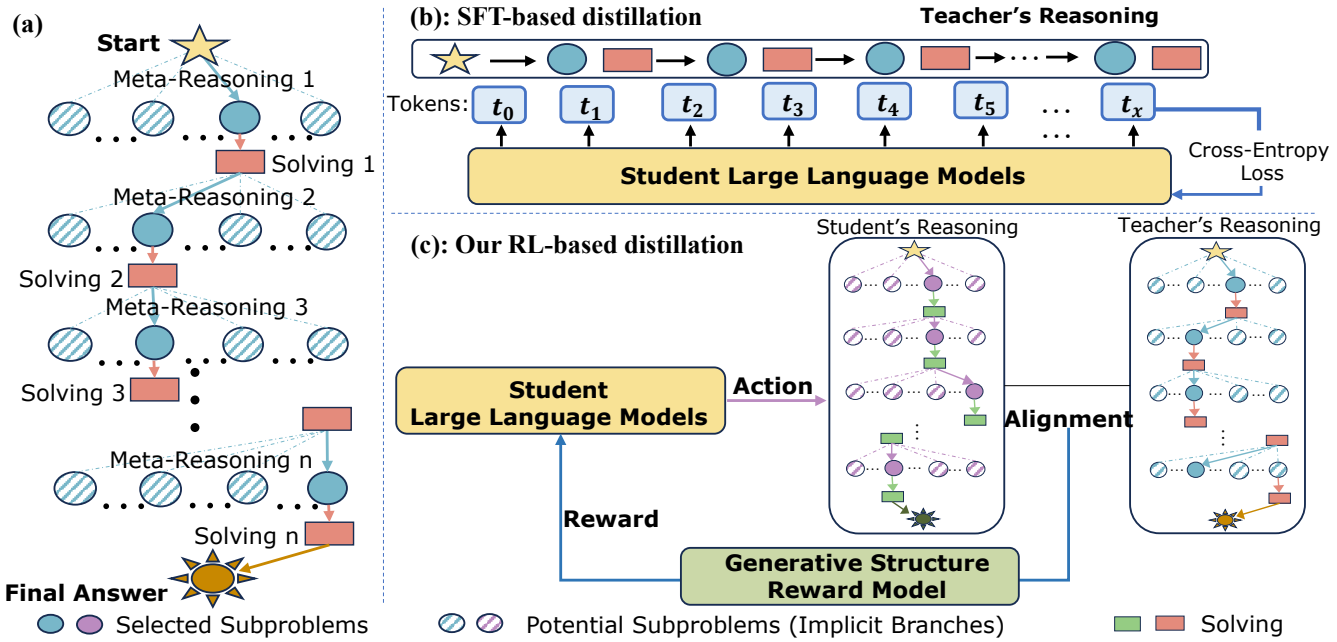


Figure 1: (a) The generated reasoning path has implicit multi-branch structure. (b) Distillation only based on SFT collapses the rich structure into a flat sequence of token prediction to memorize only the teacher's generated path. (c) Our proposed RL-based distillation can teach the student LLM to learn this structure by using a Generative Structure Reward Model to measure the alignment between the reasoning structure of the student and teacher, serving as the reward in RL.

tive Structure Reward Model (GSRM), a two-stage reward pattern that combines the semantic understanding of generative reward model (Mahan et al. 2024) with the interpretability and controllability of the rule-based reward model (Guo et al. 2025). GSRM can convert the reasoning path into the sequence consisting of multiple meta-reasoning-solving steps, and then score the matching degree between the sequences of the student and the teacher with structured reward mechanism. When combing GSRM with RL, our RLKD can guide the student LLM at a step-level on how to better perform sampling to select the most suitable sub-problem from multiple potential ones , and then solving it. In this way, RLKD can distill the implicit multi-branch structure in the reasoning of teacher LLMs to the student LLMs (Figure 1 (c)).

Experiments on math and graduate-Level Q&A show that our RLKD: (1) can use only 0.1% training data with RL-only paradigm to outperform SFT-RL pipeline in Qwen2.5-Math, (2) can further unleash the potential reasoning ability of the student LLM than SFT-based distillation, and (3) can outperform existing RL baselines. In conclusion, this paper points out that the authentic reasoning possesses an implicit multi-branch structure, which can not be distilled to the student LLM by SFT and proposes RLKD, which is the first RL-based distillation method for LLM's reasoning.

2 Related Work

2.1 Structure in Reasoning of LLMs

Recent studies finds that the reasoning of LLMs has the structure. Chain-of-Thought (Wei et al. 2022), Least-to-Most (Zhou et al. 2022) and Self-Ask (Press et al. 2022) initially formalize the reasoning path as a chain consisting of multiple nodes, which is a linear structure. Tree-of-Thought (Yao et al. 2023), Graph-of-Thought (Besta et al. 2024) and SearChain (Xu et al. 2024a) explicitly builds a non-linear structure for reasoning. SuperCorrect (Yang et al. 2025) uses high-level plans plus detailed steps as hierarchical thought templates to correct student models. The commonality of existing studies is that they explicitly let LLMs generated the specified structure in reasoning through prompt engineering or fine-tuning. We rethink the reasoning that a generated reasoning path contains multiple potential other paths, and these potential paths together with the generated path form the structure of reasoning. We focus on distilling this implicit multi-branch structure into the student LLM.

2.2 SFT for Reasoning Distillation

Supervised fine-tuning (SFT) on chain-of-thought demonstrations has emerged as a straightforward way to distill reasoning capabilities from large models into smaller ones. For example, Deepseek releases a series distilled LLMs based on Deepseek-R1 and significantly improve the reasoning capabilities (Guo et al. 2025). A similar study shows that with only 17k curated reasoning traces, a 32B student model can nearly match a closed-source o1-preview on math and coding benchmarks (Li et al. 2025) and many open-source projects are released (Wen et al. 2025; Face 2025; Bespoke-Labs 2025). However, recent findings highlight that SFT often teaches format over substance: models learn to imitate the reasoning paths without authentic understanding of their content (Chen et al. 2025; Dai et al. 2024). In fact, a student can produce cor-

rect answers by mimicking a long Chain-of-Thought (CoT) pattern even if many intermediate steps are incorrect (Li et al. 2025). The key reason behind this is that reasoning has implicit multi-branch structure but SFT distillation collapses this structure into a flat sequence of token prediction. Our RL-based distillation can teach the student LLM to learn this.

2.3 Reinforcement Learning for LLMs' Reasoning

Reinforcement learning (RL) has been explored as a means to optimize reasoning strategies in language models, building on foundations like Proximal Policy Optimization (PPO) (Schulman et al. 2017). More recent advances include Group Relative Policy Optimization (GRPO) (Shao et al. 2024), introduced by the DeepSeek team to push mathematical reasoning performance and Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) (Yu et al. 2025) to train LLM with RL at Scale. These methods primarily focus on human feedback, outcome accuracy or heuristic rewards to optimize LLMs in specific downstream tasks. Different from them, our RLKD aims to use RL in knowledge distillation.

3 Our Method

This section introduces details of our RLKD, a RL-based knowledge distillation method that can transfer the implicit multi-branch structure for complex reasoning from teacher to student. Firstly, we propose to train a Generative Structure Reward Model (GSRM) to score the alignment degree between the reasoning paths of student and teacher LLMs in terms of their implicit multi-branch structure. Then, we introduce the GSRM into RL-based knowledge distillation.

3.1 Generative Structure Reward Model

Rewarding the implicit multi-branch structure is premised on accessing the meta-reasoning and solving at each reasoning step. However, the raw reasoning path generated by LLMs such as Deepseek-R1 is unstructured, making it difficult to directly distinguish between various reasoning steps and to decouple the meta-reasoning and solving content of each step. To solve this, we propose Generative Structure Reward Model (GSRM), a two-stage reward pattern that combines the semantic understanding of generative reward model (Mahan et al. 2024) with the interpretability and controllability of the rule-based reward model (Guo et al. 2025). In the implementation of GSRM, firstly, we train a LLM to generate a sequence of meta-reasoning and solving pairs for the input reasoning path. Then, we design a structured reward mechanism to score the alignment degree between the teacher and the student in terms of their meta-reasoning and solving content at the corresponding steps.

Dataset Construction. We devise detailed instructions and examples to perform in-context learning (ICL) with the GPT-4o API, enabling it to automatically construct a large-scale supervised fine-tuning dataset. Each data sample of this dataset is a input-out pair: reasoning path $\mathbf{R}$ is the input and the sequence $\mathbf{S}$ consisting of multiple meta-reasoning and solving steps is the output. Specifically, we first structurally define each meta-reasoning and solving steps as ($\mathcal{M}$, $\mathcal{Q}$ & $\mathcal{A}$), in which $\mathcal{M}$ is the content of meta-reasoning that focuses on determining the sub-problem that the current reasoning step should solve. $\mathcal{Q}$-$\mathcal{A}$ pair is the content of solving, in which $\mathcal{Q}$ is a clear description of the current sub-problem and $\mathcal{A}$ is the solving result to the sub-problem. So sequence $\mathbf{S}$ containing n meta-reasoning and solving steps can be described as:

$$\mathbf{S} = [(\mathcal{M}_1, \mathcal{Q}_1 \& \mathcal{A}_1), (\mathcal{M}_2, \mathcal{Q}_2 \& \mathcal{A}_2), ..., (\mathcal{M}_n, \mathcal{Q}_n \& \mathcal{A}_n)].$$

The essential for the generation of $\mathbf{S}$ is (1) containing each key reasoning step in the reasoning path and (2) fully decoupling the contents of $\mathcal{M}$, $\mathcal{Q}$, and $\mathcal{A}$ so that they contain and only contain the specified information. Based on these two points, we design specific instruction and examples covering four reasoning tasks (math, science, code and puzzles) to enable GPT-4o to perform effective ICL. One example is shown in Figure 2 (b). Besides, we introduce a verification-based feedback strategy to improve data quality. For each $\mathbf{S}$ generated by GPT-4o given input $\mathbf{R}$, we use Deepseek-V3 to determine whether $\mathbf{S}$ meets the requirements of the instruction. If it does not, we give the feedback from Deepseek-V3 to GPT-4o to re-generate $\mathbf{S}$. If a data sample fails to pass the verification after three re-generations, we discard this sample. We automatically execute this data production process on OpenThoughts-114k (Team 2025), an open synthetic reasoning dataset with 114k high-quality examples covering math, science, code, and puzzles, and each example has Deepseek-R1 generated reasoning path, which is the $\mathbf{R}$. Finally, we get $93,625$ $\mathbf{R}$-$\mathbf{S}$ pairs for the future supervised fine-tuning.

Structured Fine-grained Training. This stage aims to train a generative reward model in our GSRM that can generate the sequence of meta-reasoning and solving pairs for the input reasoning path. Since the target output $\mathbf{S}$ is a highly structured text, we propose a training method called Structured Fine-grained Training to optimize each task (meta-reasoning generation and solving generation) in a fine-grained manner and dynamically adjust the optimization weight according to the difficulty of the task. Specifically, we split the tokens in $\mathbf{S}$ into two sets: one are tokens in meta-reasoning $\mathcal{M}$ and the other are tokens in solving $\mathcal{Q}\&\mathcal{A}$. We perform 3 training epochs. $\mathcal{F}$ is the LLM used in this part (Qwen2.5-7B-Instruct), the first epoch is training for meta-reasoning, only loss from the tokens in $\mathcal{M}$ is considered:

$$\mathcal{L}_1 = \sum_{\mathbf{S}_{[i]} \in \mathcal{M}} -\log \mathcal{F}(\mathbf{S}_{[i]} | \mathbf{S}_{[1:i-1]}; \theta).$$

The second epoch is training for solving generation, only loss from $\mathcal{Q}\&\mathcal{A}$ is computed as:

$$\mathcal{L}_2 = \sum_{\mathbf{S}_{[i]} \in \mathcal{Q}\&\mathcal{A}} -\log \mathcal{F}(\mathbf{S}_{[i]} | \mathbf{S}_{[1:i-1]}; \theta).$$

The third epoch is mixed training. Total loss is the dynamically weighted average of the losses on the two token sets:

$$\begin{aligned} \mathcal{L}_3 = a \sum_{\mathbf{S}_{[i]} \in \mathcal{M}} -\log \mathcal{F}(\mathbf{S}_{[i]} | \mathbf{S}_{[1:i-1]}; \theta) \\ + b \sum_{\mathbf{S}_{[i]} \in \mathcal{Q}\&\mathcal{A}} -\log \mathcal{F}(\mathbf{S}_{[i]} | \mathbf{S}_{[1:i-1]}; \theta). \end{aligned} \tag{1}$$

[Question]: Given real numbers a, b, c and a positive number λ such that the polynomial $f(x) = x^3 + ax^2 + bx + c$ has three real roots x_1, x_2, x_3, and the conditions $x_2 - x_1 = \lambda$ and $x_3 > \frac{1}{2}(x_1 + x_2)$ are satisfied, find the maximum value of $\frac{2a^3+27c-9ab}{\lambda^3}$.
[Reasoning Path]: Okay, so I... (too long to show here)
[Transformed Response]:
[Meta-reasoning 1]: To find the maximum value of $\frac{2a^3+27c-9ab}{\lambda^3}$, use Vieta's formulas to express coefficients a, b, c in terms of the roots x_1, x_2, x_3. Parametrize the roots to satisfy $x_2 - x_1 = \lambda$ and $x_3 > \frac{1}{2}(x_1 + x_2)$. Define $x_1 = t - \frac{\lambda}{2}$, $x_2 = t + \frac{\lambda}{2}$, and $x_3 = t + s$ where $s > 0$. This parametrization simplifies the problem by eliminating λ through symmetry.
[Sub-problem 1]: Express a, b, b in terms of t, s, and λ using the parametrized roots. [Answer 1]: $a = -3t - s, b = 3t^2 + 2ts - \frac{\lambda^2}{4}, c = -t^3 - t^2 s + \frac{\lambda^2}{4}(t + s)$
[Meta-reasoning 2]: Substitute a, b, c into the expression $\frac{2a^3+27c-9ab}{\lambda^3}$. Simplify the expression by combining like terms. The result will depend only on s and λ, allowing further simplification by introducing $k = \frac{s}{\lambda}$. This reduces the problem to maximizing a cubic function in k.
[Sub-problem 2]: Simplify $\frac{2a^3+27c-9ab}{\lambda^3}$ after substitution. What is the simplified form in terms of k? [Answer 2]: $-2k^3 + \frac{9}{2}k$.
[Meta-reasoning 3]: To maximize $Q(k) = -2k^3 + \frac{9}{2}k$, find its critical points by taking the derivative and solving for k. Verify that the critical point corresponds to a maximum using the second derivative test.-
[Sub-problem 3]: Find the critical point of $Q(k) = -2k^3 + \frac{9}{2}k$. [Answer 3]: $k = \frac{\sqrt{3}}{2}$.
[Meta-reasoning 4]: Substitute $k = \frac{\sqrt{3}}{2}$ back into $Q(k)$ to compute the maximum value. Confirm the calculation to ensure no arithmetic errors.-
[Sub-problem 4]: Calculate $Q(\frac{\sqrt{3}}{2})$. [Answer 4]: $\frac{3\sqrt{3}}{2}$.
[Meta-reasoning 5]: Verify that the critical point $k = \frac{\sqrt{3}}{2}$ satisfies all constraints (e.g., $s > 0$) and that the polynomial discriminant condition for three real roots is implicitly satisfied by the problem's given conditions. Confirm no additional constraints affect the result.-
[Sub-problem 5]: Does $k = \frac{\sqrt{3}}{2}$ satisfy $s > 0$? [Answer 5]: Yes.
[Meta-reasoning 6]: Finalize the maximum value of the original expression based on the derived result.-
[Sub-problem 6]: What is the maximum value of $\frac{2a^3+27c-9ab}{\lambda^3}$? [Answer 6]: $\frac{3\sqrt{3}}{2}$.
So the final answer is: $\frac{3\sqrt{3}}{2}$.

Figure 2: One sequence generation example for math task in in-context learning prompts for GPT-4o.

a and b are initialized to the average of $\mathcal{L}_1$ and $\mathcal{L}_2$ over training steps respectively. During training, we calculate the loss of the tokens in $\mathcal{M}$ and $\mathcal{Q}\&\mathcal{A}$ respectively, and update a and b according to the ratio of the losses in two sets every α training steps. We set α as 100.

Structured Reward Mechanism After above generation, we denote $\mathbf{S^t}$ and $\mathbf{S^s}$ as the sequences generated from reasoning paths of the teacher and the student LLM respectively:

$$\mathbf{S^t} = [(\mathcal{M}_1^t, \mathcal{Q}_1^t \& \mathcal{A}_1^t), (\mathcal{M}_2^t, \mathcal{Q}_2^t \& \mathcal{A}_2^t), ..., (\mathcal{M}_n^t, \mathcal{Q}_n^t \& \mathcal{A}_n^t)],$$
$$\mathbf{S^s} = [(\mathcal{M}_1^s, \mathcal{Q}_1^s \& \mathcal{A}_1^s), (\mathcal{M}_2^s, \mathcal{Q}_2^s \& \mathcal{A}_2^s), ..., (\mathcal{M}_m^s, \mathcal{Q}_m^s \& \mathcal{A}_m^s)].$$

We propose the Structured Reward Mechanism to map the generated sequence consisting of multiple meta-reasoning and solving to a reward value according to the alignment between $\mathbf{S^t}$ and $\mathbf{S^s}$. Although this computation is carried out based on a linear sequence, when combined with RL, it becomes capable of quantifying the degree of alignment between the teacher and the student in terms of the implicit multi-branch structure, serving as environmental feedback in RL. This is because it can assess the step-level alignment between the sequences and, through the mechanism of reward for RL, finely guide the student LLM to make correct sampling from multiple potential sub-problems at each meta-reasoning step. Compared with SFT-based distillation, it makes the student LLM learn the implicit multi-branch structure in the authentic reasoning of the teacher LLM rather than just focuse on memorizing the reasoning path on the surface generated by the teacher LLM. Compared with existing RL methods, it can guide the student LLM on how to do better sampling at each step and avoid reward hacking (Amodei et al. 2016; Di Langosco et al. 2022) by step-to-step comparison and early-exist mechanism.

Algorithm 1: Structured Reward Mechanism

```
Initialize r ← 0 ;  // Total reward accumulator
for each index i ∈ {1, 2, ..., min(n, m)} do
    Initialize temporary reward v ← 0 ;  // Step-wise reward score
    if Match(M_i^t, M_i^s) then
        v ← 1 ;      // Base score for matched meta-reasoning
        if Not Match(Q_i^t, Q_i^s) then
            v ← v × 0.5 ;      // 50% penalty for question mismatch
        end
        if Not Match(A_i^t, A_i^s) then
            v ← v × 0.5 ;      // Additional 50% penalty for answer mismatch
        end
    end
    else
        Break ;     // Encountering mismatched meta-reasoning, exit
    end
    r ← r + v ;              // Accumulate step contribution
end
return r ; // Final reward between sequences
```

Model	Data Size		AIME24										
	SFT	RL	pass@1						pass@4	pass@8	pass@16	pass@32	pass@64
Number of running times for each question			1	4	8	16	32	64	4	8	16	32	64
Based on Qwen2.5-Math-7B													
Qwen2.5-Math-7B	0	0	13.3	11.7	11.7	11.3	11.0	10.3	20.0	26.7	30.0	30.0	43.3
Qwen2.5-Math-7B-Instruct	2,895K	66K	16.7	15.8	15.8	15.8	14.7	14.6	33.3	43.3	43.3	50.0	50.0
Our: Qwen2.5-Math-7B-RLKD-Zero	0	3.2K	23.3	20.0	20.4	22.1	22.5	21.0	40.0	46.7	56.7	60.0	70.0
Based on Deepseek-R1-Distill-Qwen-7B													
DeepSeek-R1-Distill-Qwen-7B	800K	0	50.0	52.7	52.5	52.9	52.3	52.4	66.7	73.3	80.0	80.0	83.3
DeepSeek-R1-Distill-Qwen-7B-PPO	800K	3.2K	46.7	52.1	53.0	52.7	52.9	52.9	66.7	73.3	73.3	80.0	83.3
DeepSeek-R1-Distill-Qwen-7B-GRPO	800K	3.2K	50.0	52.5	53.3	**53.3**	53.3	52.3	66.7	73.3	80.0	83.3	83.3
Our: DeepSeek-R1-Distill-Qwen-7B-RLKD	800K	3.2K	**53.3**	**56.7**	**55.4**	**53.3**	**52.9**	**53.6**	**73.3**	**80.0**	**86.7**	**86.7**	**86.7**

(a) Performance on AIME24

Model	Data Size		GPQA-Diamond					MATH-500		
	SFT	RL	pass@1			pass@4	pass@8	pass@1		pass@4
Number of running times for each question			1	4	8	4	8	1	4	4
Based on Qwen2.5-Math-7B										
Qwen2.5-Math-7B	0	0	29.3	27.8	27.5	61.1	79.3	54.8	56.2	81.0
Qwen2.5-Math-7B-Instruct	2,895K	66K	30.3	30.2	30.1	66.7	82.3	82.4*	81.5*	89.4*
Our: Qwen2.5-Math-7B-RLKD-Zero	0	3.2K	34.9	34.2	32.7	69.2	86.4	74.4	73.9	87.8
Based on Deepseek-R1-Distill-Qwen-7B										
DeepSeek-R1-Distill-Qwen-7B	800K	0	47.9	50.8	50.2	74.7	83.8	92.4	93.2	97.4
DeepSeek-R1-Distill-Qwen-7B-PPO	800K	3.2K	47.9	49.4	50.7	74.7	84.3	93.4	94.0	97.6
DeepSeek-R1-Distill-Qwen-7B-GRPO	800K	3.2K	50.5	50.1	49.8	75.3	84.3	93.0	93.7	97.4
Our: DeepSeek-R1-Distill-Qwen-7B-RLKD	800K	3.2K	**54.5**	**53.0**	**52.8**	**76.8**	**86.9**	**94.2**	**95.1**	**98.2**

(b) Performance on GPQA-Diamond and MATH-500

*means training set of MATH-500 has appeared in the SFT training data of Qwen2.5-Math-7B-Instruct (Yang et al. 2024).

Table 1: Reasoning abilities on AIME24, MATH-500 and GPQA-Diamond. The results are obtained based on generating multiple responses for each query to mitigate randomness and the best results are in **bold** font. Qwen2.5-Math-7B-RLKD-Zero is trained by our RLKD without any SFT (RL only).

Specifically, our structured reward mechanism sequentially compares the corresponding steps of $\mathbf{S^t}$ and $\mathbf{S^s}$ (Algorithm 1). We use Qwen-2.5-7B-Instruct to determine whether two texts are matched. For the steps where $\mathcal{M}_i^t$ matches $\mathcal{M}_i^s$, we assign a temporary reward value of 1. We further deduct the temporary reward by judging the matching relationship of $\mathcal{Q}_i$ and $\mathcal{A}_i$. The temporary reward value is added to the total reward value after this. When meeting the mismatched $\mathcal{M}_i$, the reward accumulation ends. This reward mechanism follows the sequential dependency of the reasoning path and can give an approximately unique reward value for each step.

3.2 RL-based Knowledge Distillation Training

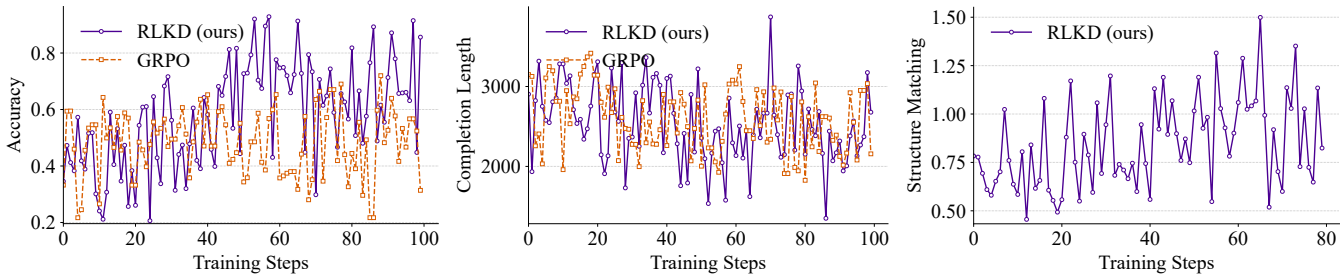
We combine our Generative Structure Reward Models (GSRM) with Group-based Relative Policy Optimization (GRPO) (Shao et al. 2024) for RL-based knowledge distillation training. In training, we combine the reward obtained from GSRM and the outcome reward of the specific task, such as the accuracy in math, in a weighted manner as the total reward for GRPO.

4 Experiments

4.1 Experimental Setup.

Datasets and Evaluation Metrics. We use OpenR1-Math as the training dataset for RL, in which Deepseek-R1 generated responses in this dataset are used as the teacher LLM's reasoning paths. We keep the training datasets consistent with baselines including PPO and GRPO. In evaluation, we use the popular and challenging datasets on LLM's reasoning including AIME24 (MAA 2024) and MATH-500 (Hendrycks et al. 2021) for math reasoning and GPQA-Diamond (Rein et al. 2024) for graduate-Level Q&A. As for the metrics, we follow existing studies on LLM's reasoning (Guo et al. 2025; Shao et al. 2024; Zhang et al. 2025; Deng et al. 2024; Xu et al. 2024b) to use pass@k (Chen et al. 2021). Our validation is divided into two parts: pass@1 and pass@k ($k > 1$). As for pass@1, we generate m (m is 64 for AIME, 8 for GPQA and 4 for MATH500) responses for each question and compute pass@1 as $pass@1 = \frac{1}{m}\sum_{i=1}^{m} p_i$, in which p_i is the correctness of the i-th response. This can alleviate randomness on small datasets. As for pass@k ($k > 1$), it involves the LLM generating k responses for each question, with the data sample being marked as accurate if at least one of the k responses is accurate. Paying attention to the metric where $k > 1$ is crucial, as in this setting, the LLM has the opportunity to explore multiple diverse paths to answer the question, it reflects the LLM's ability to sample from multiple implicit paths during reasoning, thereby assessing whether the distilled LLM has learned authentic reasoning or merely memorized the teacher's paths.

Baselines and Comparison Settings. We categorize the baselines into three groups according to the experimental settings. The first setting aims to show the effect of our method on improving LLM's reasoning ability. We compare our RL-only training method with the LLM trained in



(a) Accuracy (b) Completion Length (c) Alignment of the Structures

Figure 3: The variations of metrics during the RL training process ((c) is not applicable to GRPO).

SFT-RL pipeline. We use Qwen2.5-Math-7B-Instruct (Yang et al. 2024), a powerful reasoning LLM trained on large-scale chain-of-thought datasets with SFT and then GRPO. Both our method and Qwen2.5-Math-7B-Instruct use the same based model: Qwen2.5-Math-7B, which has been pre-trained on math corpus. The second setting aims to explore whether our RL-based distillation can further improve the performance of the SFT-distilled LLM and achieve effective optimization of the SFT-RL pipeline in knowledge distillation. We compare our method with RL baselines including PPO and GRPO based on DeepSeek-R1-Distill-Qwen-7B (Guo et al. 2025), a powerful LLM that is SFT-distilled from Deepseek-R1. The third setting aims to compare which method is better for the student LLM to learn authentic reasoning rather than memorizing the teacher's path. We use embedding[1] similarity (Deng et al. 2025) to select the 3.2K data samples with the largest difference from the test set in OpenR1-Math-220k training set, and train Qwen2.5-Math-7B on this subset with RL-based distillation and SFT-based distillation respectively.

Implementation Details. In training, we build our code based on Open-R1, an open-source project for LLM's reasoning. We use Pytorch 2.5.1 as the training framework and Deepspeed 0.15.4 for acceleration of parallel computing. In RL training, we use one Ascend 910B 64G NPU for online inference and four Ascend 910B 64G NPUs for training under deepspeed-zero3 setting. As for hyperparameters, we set per-device batch size as 2, gradient accumulation steps as 4, group size for GRPO as 4, temperature in online inference as 0.7. In evaluation, we use Lighteval as the toolkit and follow settings in Deepseek (Guo et al. 2025) to set temperature as 0.6, max new tokens as 32768 and top-p as 0.95. We report the results of multiple runs to reduce the randomness.

4.2 Experimental Results

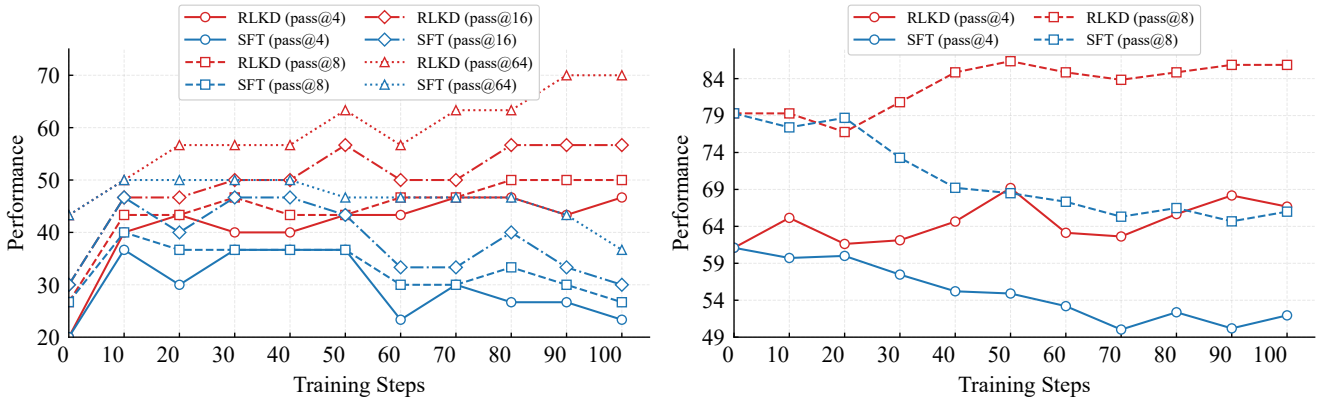
Main Results. Results about reasoning abilities of LLMs are shown in Table 1. In the training based on Qwen2.5-Math-7B, our RL-only method Qwen2.5-Math-7B-RLKD-Zero outperforms complex SFT + RL pipeline (Qwen2.5-Math-7B-Instruct) and uses much less data (nearly 0.1%). In the training based on Deepseek-R1-Distill-Qwen-7B (SFT-distilled LLM), baseline RL methods including PPO and GRPO can hardly bring about significant improvements on this basis while our RLKD is capable of further enhancing performance. The indicates that our RL-based distillation approach enables the SFT-distilled LLM to learn additional information beyond its memorization of the teacher's reasoning paths. The relatively more significant improvements are observed in pass@k ($k > 1$). In this setting, LLM has the opportunity to explore multiple diverse paths to answer the question, which means that compared to SFT distillation, our method enables the student LLM to learn how to perform sampling from multiple potential paths by distilling the implicit multi-branch structure from the teacher, thereby increasing the probability of providing the correct answer.

Ablation Study. Figure 3 shows three metrics in RL training. We compare our RLKD with GRPO because RLKD is actually GRPO with the reward from our Generative Structure Reward Model (GSRM). The results indicate that as the training progresses, GSRM enables RLKD to better optimize the accuracy of the task (Figure 3 (a)), primarily because the student gradually learns the teacher's implicit multi-branch structure in reasoning (Figure 3 (c)).

4.3 Compare RL Distillation with SFT Distillation

Performance Trend Varying with Training Steps. This section compares the SFT-based distillation with our RL-based distillation (RLKD) by training Qwen2.5-Math-7B on the dataset has domain shift to AIME24 and is out of the domain of GPQA. This setting, where there exists a domain discrepancy between the training and testing sets for distillation, allows us to intuitively discern whether SFT-based distillation is merely mimicking and memorizing the teacher's paths, and whether our RL-based distillation enables the student to learn authentic reasoning. The experimental results are shown in Figure 4. It is noteworthy that, as the training progresses, SFT and RLKD demonstrate completely opposite performance trends: RLKD can consistently enhance performance, even when distilling on dataset that significantly diverges from the testing set, whereas SFT progressively undermines performance. It indicates that SFT distillation is easy to fall into the trap of simply imitating and memorizing the teacher's reasoning paths, rather than learning authentic reasoning that can be stably generalized. Our RLKD teaches students to sample from multiple potential branches, much like the multi-branch structure implicit in real reasoning in teacher LLMs.

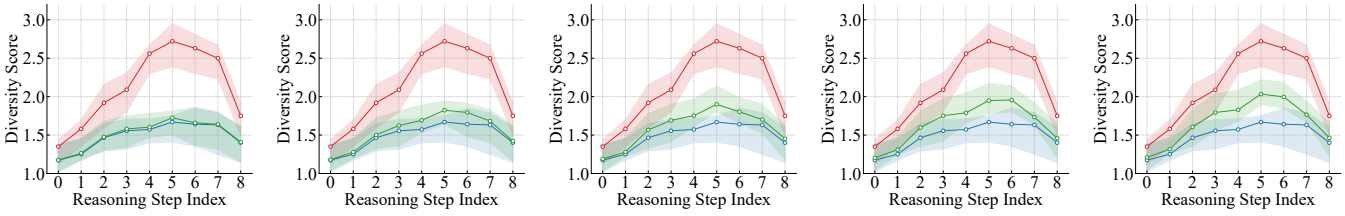
[1] obtained by gte-Qwen2-7B-instruct



(a) AIME24 (Domain Shift)

(b) GPQA-Diamond (Out-of-Domain)

Figure 4: Comparison between SFT-based distillation and our RL-based distillation (RLKD) in domain shift and out-of-domain setting. SFT and RLKD see the same data (32 samples) at each step.



(a) Training Step 20 (b) Training Step 40 (c) Training Step 60 (d) Training Step 80 (e) Training Step 100

Figure 5: Diversity among different reasoning paths at each meta-reasoning step varying with training. Red line is Deepseek-R1. Blue line is DeepSeek-R1-Distill-Qwen-7B. Green line is DeepSeek-R1-Distill-Qwen-7B w/ RLKD.

Diversity of Reasoning Paths. The diversity of reasoning paths can reflect whether the student LLM has truly learned reasoning or is just memorizing the fixed teacher's path in the training set. Figure 5 shows that our method brings the diversity-patterns of the student LLM closer to that of the teacher LLM in multi-step reasoning, suggesting that it allows the student LLM to learn authentic reasoning by distilling the implicit multi-branch structure. Specifically, we randomly sample 500 samples from Openthoughts-114K. On this sampled set, we set temperature to 0.8 and let LLM generate 16 responses for each question. We use the GSRM in Section 3.1 to generate a corresponding sequence containing multiple meta-reasoning-solving steps for each response, and calculate the diversity of the meta-reasoning-solving content within the 16 responses of each question at step-level. The diversity score D is calculated as:

$$\mathbf{u} = \frac{1}{16}\sum_{i=1}^{16}\mathbf{e}_i, \quad D = \frac{1}{\frac{1}{16}\sum_{i=1}^{16}\frac{\mathbf{e}_i \cdot \mathbf{u}}{|\mathbf{e}_i||\mathbf{u}|}},$$

in which $\mathbf{e}_i$ is the text embedding encoded by *gte-Qwen2-7B-instruct* for each meta-reasoning-solving content. As shown in Figure 5, the teacher LLM (Deepseek-R1) has a significantly different diversity-pattern from the SFT-distilled student LLM (DeepSeek-R1-Distill-Qwen-7B). Reasoning in the teacher has the higher diversity while the student is stuck in the relatively fixed paths. As the training of our method progresses (from step 20 to 100), the diversity of the student's reasoning paths begins to increase and gradually approaches that of the teacher, which indicates that our method allows the student to learn the teacher's authentic reasoning paradigm.

5 Conclusion and Discussion

This work addresses a critical flaw in knowledge distillation in LLM's reasoning: the failure of SFT to transfer the implicit multi-branch structure underlying authentic reasoning. Drawing from cognitive neuroscience principles, we show that authentic reasoning involves dynamic meta-reasoning (sub-problem selection) and solving steps—a implicit multi-branch structure flattened by token-level SFT training. Our RLKD, the first RL-based distillation framework for reasoning, overcomes this paired with a Generative Structure Reward Model (GSRM), which decomposes reasoning paths into meta-reasoning-solving pairs and scores the structural alignment between teacher and student. Experiments across math and graduate-level QA tasks demonstrate RLKD's superiority than SFT-based distillation, SFT-RL pipeline and RL baselines including PPO and GRPO, proving its ability to distill how teachers navigate latent reasoning branches rather than mimicking surface tokens. Further analysis confirms RLKD-trained students mirror teachers' multi-branch exploration patterns, closing the imitation-authentic reasoning gap.

Acknowledgments

This work was supported by the Key Research and Development Program of Xinjiang Uyghur Autonomous Region Grant No. 2024B03026, the Strategic Priority Research Program of the CAS under Grants No.XDB0680302, the Beijing Nova Program under Grants No. 20250484765, the National Natural Science Foundation of China (NSFC) under Grants No. 62276248, and the Youth Innovation Promotion Association CAS under Grants No. 2023111.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Bespoke-Labs. 2025. Bespoke-stratos: The unreasonable effectiveness of reasoning distillation., January 2025. https://huggingface.co/bespokelabs/Bespoke-Stratos-7B. Accessed: 2025-04-07.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17682–17690.

Chen, H.; Tu, H.; Wang, F.; Liu, H.; Tang, X.; Du, X.; Zhou, Y.; and Xie, C. 2025. SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models. *arXiv preprint arXiv:2504.11468*.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Cox, M.; and Raja, A. 2007. Metareasoning: A manifesto. *BBN Technical*.

Dai, C.; Li, K.; Zhou, W.; and Hu, S. 2024. Beyond imitation: Learning key reasoning steps from dual chain-of-thoughts in reasoning distillation. *arXiv preprint arXiv:2405.19737*.

Deng, J.; Jiang, Z.; Pang, L.; Chen, L.; Xu, K.; Wei, Z.; Shen, H.; and Cheng, X. 2025. Following the Autoregressive Nature of LLM Embeddings via Compression and Alignment. *arXiv preprint arXiv:2502.11401*.

Deng, J.; Wei, Z.; Pang, L.; Ding, H.; Shen, H.; and Cheng, X. 2024. Everything is Editable: Extend Knowledge Editing to Unstructured Data in Large Language Models. *arXiv preprint arXiv:2405.15349*.

Di Langosco, L. L.; Koch, J.; Sharkey, L. D.; Pfau, J.; and Krueger, D. 2022. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, 12004–12019. PMLR.

Face, H. 2025. Open r1: A fully open reproduction of deepseek-r1, January 2025. https://github.com/huggingface/open-r1. Accessed: 2025-05-26.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Havrilla, A.; Du, Y.; Raparthy, S. C.; Nalmpantis, C.; Dwivedi-Yu, J.; Zhuravinskyi, M.; Hambro, E.; Sukhbaatar, S.; and Raileanu, R. 2024. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*.

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

Jiaqi, W.; Xinliang, L.; Zhengliang, L.; Zihao, W.; Tianyang, Z.; Peng, S.; Yiwei, L.; Hanqi, J.; Yifan, Z.; Junhao, C.; et al. 2025. LLM Reasoning: from OpenAI O1 to DeepSeek R1.

Li, D.; Cao, S.; Griggs, T.; Liu, S.; Mo, X.; Tang, E.; Hegde, S.; Hakhamaneshi, K.; Patil, S. G.; Zaharia, M.; et al. 2025. LLMs Can Easily Learn to Reason from Demonstrations Structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*.

MAA. 2024. American invitational mathematics examination - aime, February 2024. In American Invitational Mathematics Examination - AIME 2024. Accessed: 2024-06-01.

Mahan, D.; Van Phung, D.; Rafailov, R.; Blagden, C.; Lile, N.; Castricato, L.; Fränken, J.-P.; Finn, C.; and Albalak, A. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pan, Q.; Ji, W.; Ding, Y.; Li, J.; Chen, S.; Wang, J.; Zhou, J.; Chen, Q.; Zhang, M.; Wu, Y.; et al. 2025. A Survey of Slow Thinking-based Reasoning LLMs using Reinforced Learning and Inference-time Scaling Law. *arXiv preprint arXiv:2505.02665*.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N. A.; and Lewis, M. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Russell, S.; and Wefald, E. 1991. Principles of metareasoning. *Artificial intelligence*, 49(1-3): 361–395.

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

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Team, O. 2025. Open Thoughts. https://open-thoughts.ai.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wen, L.; Cai, Y.; Xiao, F.; He, X.; An, Q.; Duan, Z.; Du, Y.; Liu, J.; Tang, L.; Lv, X.; et al. 2025. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *arXiv preprint arXiv:2503.10460*.

Xu, F.; Hao, Q.; Zong, Z.; Wang, J.; Zhang, Y.; Wang, J.; Lan, X.; Gong, J.; Ouyang, T.; Meng, F.; et al. 2025a. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv preprint arXiv:2501.09686*.

Xu, S.; Pang, L.; Shen, H.; and Cheng, X. 2025b. A Theory for Token-Level Harmonization in Retrieval-Augmented Generation. In *The Thirteenth International Conference on Learning Representations*.

Xu, S.; Pang, L.; Shen, H.; Cheng, X.; and Chua, T.-S. 2024a. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In *Proceedings of the ACM Web Conference 2024*, 1362–1373.

Xu, S.; Pang, L.; Yu, M.; Meng, F.; Shen, H.; Cheng, X.; and Zhou, J. 2024b. Unsupervised Information Refinement Training of Large Language Models for Retrieval-Augmented Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 133–145.

Yang, A.; Zhang, B.; Hui, B.; Gao, B.; Yu, B.; Li, C.; Liu, D.; Tu, J.; Zhou, J.; Lin, J.; et al. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Yang, L.; Yu, Z.; Zhang, T.; Xu, M.; Gonzalez, J. E.; CUI, B.; and YAN, S. 2025. SuperCorrect: Advancing Small LLM Reasoning with Thought Template Distillation and Self-Correction. In *The Thirteenth International Conference on Learning Representations*.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Fan, T.; Liu, G.; Liu, L.; Liu, X.; et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Yue, Y.; Chen, Z.; Lu, R.; Zhao, A.; Wang, Z.; Song, S.; and Huang, G. 2025. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? *arXiv preprint arXiv:2504.13837*.

Zhang, C.; Deng, Y.; Lin, X.; Wang, B.; Ng, D.; Ye, H.; Li, X.; Xiao, Y.; Mo, Z.; Zhang, Q.; et al. 2025. 100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models. *arXiv preprint arXiv:2505.00551*.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q.; et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.