Reinforcement Learning with Verifiable Rewards Incentivizes Correct Reasoning in Base LLMs

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, Mao Yang

preprint (ICLR 2026)

Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs

ICLR 2026 Conference Submission9292 Authors

Published: 26 Jan 2026, Last Modified: 26 Jan 2026 ICLR 2026 Everyone Revisions BibTeX

2026. 02. 04

Learning Agents 강화학습 논문 리뷰 스터디

Minkyoung Kim

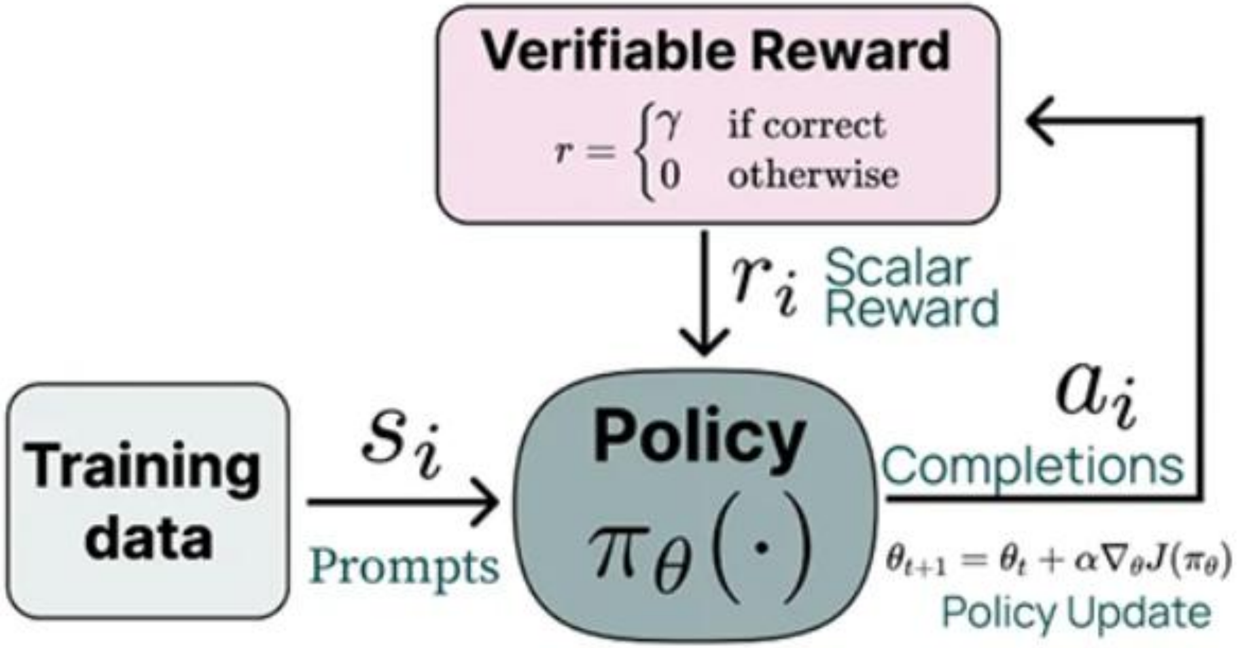
REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS IMPLICITLY INCENTIVIZES CORRECT REASONING IN BASE LLMS

55회 인용(26.02.02)

Xumeng Wen[*1], Zihan Liu[*†2], Shun Zheng[*‡1], Shengyu Ye[†1], Zhirong Wu[1], Yang Wang[1], Zhijian Xu[†3], Xiao Liang[†4], Junjie Li[1], Ziming Miao[1], Jiang Bian[1], Mao Yang[1]
[1]Microsoft Research Asia [2]Peking University
[3]The Chinese University of Hong Kong [4]University of California, Los Angeles

RLVR(Reinforcement Learning with Verifiable Rewards)

- 모델의 출력이 미리 정해진 정답 기준을 만족할 때만 보상을 주는 방법
- 명확한 정답 신호를 활용함으로써, LLM을 보다 객관적으로 신뢰할 수 있는 기준으로 훈련
- e.g. 수학 문제 or 프로그래밍 코드

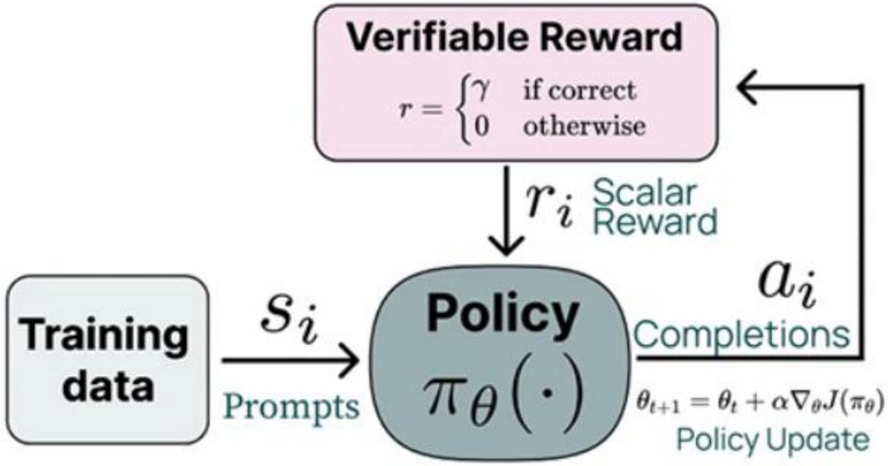


https://www.youtube.com/watch?v=skT89EvIjrc
https://www.lgresearch.ai/blog/view?seq=565
https://wikidocs.net/278478

Introduction

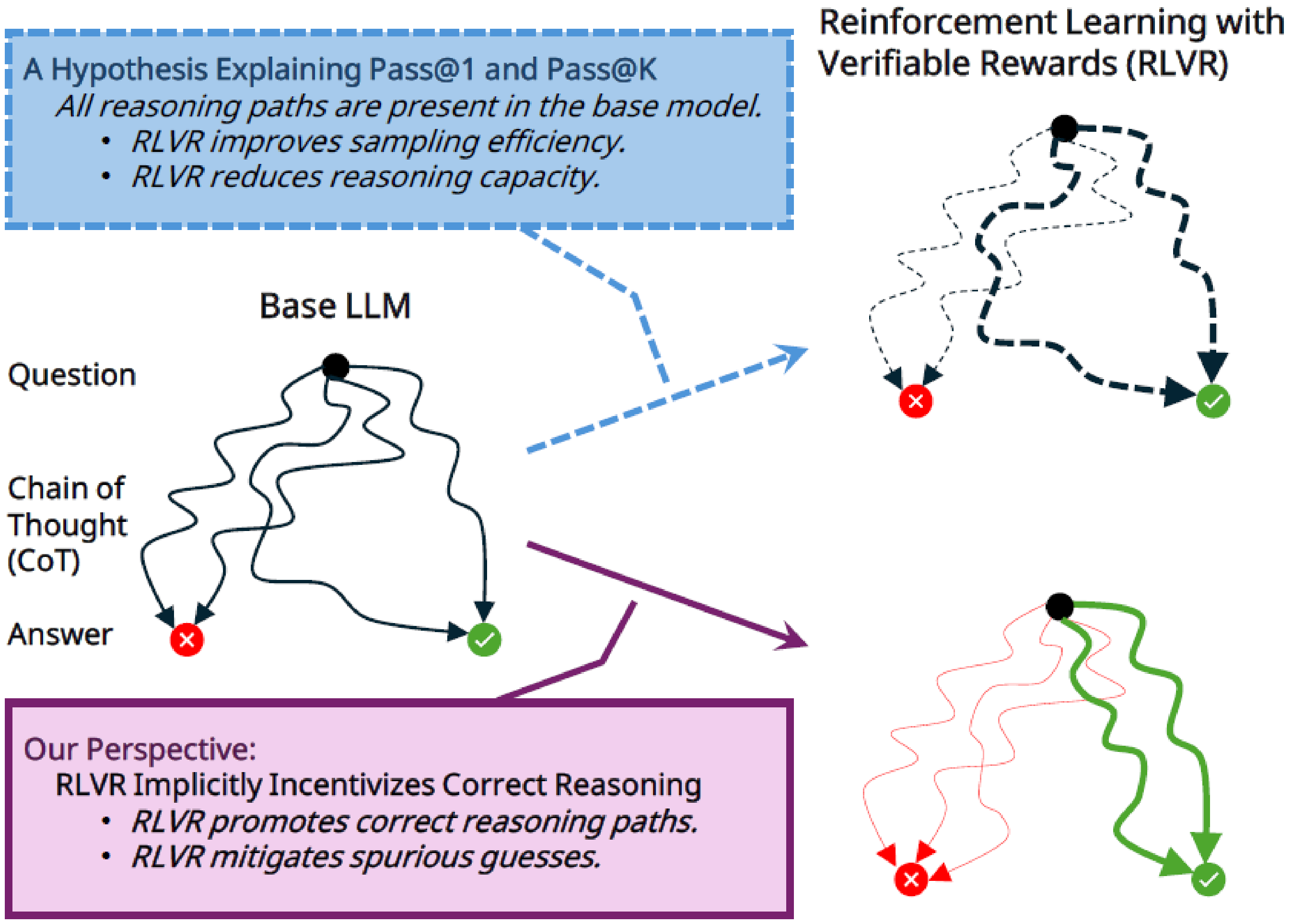
- Recent advancement in long **chain-of-thought (CoT) reasoning** have led to significant interest in potential Reinforcement Learning with Verifiable Rewards (RLVR) for Large Language Models(LLMs).
  (through the Group Relative Policy Optimization algorithm used by DeepSeek-R1)

- Reinforcement Learning with Verifiable Rewards(RLVR)
  - Large Language Model(LLM) acts a policy, generating a CoT as a sequence of actions and receiving feedback on answer correctness from **deterministic verifiers**.
  - This paradigm holds the promise of endowing LLM with the ability to learn from experience through **free exploration**.



Introduction

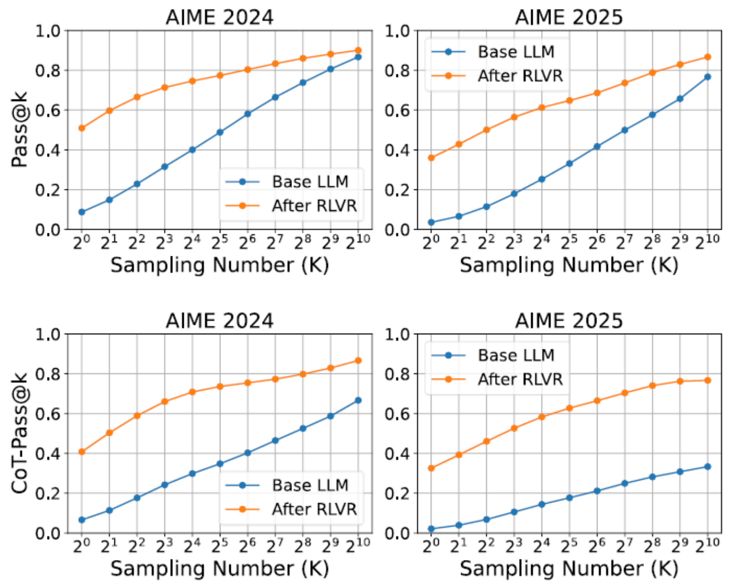
- **Debates** on Whether RLVR really Incentivizes
  - Whether it truly **enhances reasoning abilities** or **simply boosts sampling efficiency**?



Introduction

- **Debate**: Whether it truly enhances reasoning abilities or simply boosts sampling efficiency?
  - Some post-RLVR model improve the Pass@1 metric,
    but fail to enhance the Pass@K metric compared to the base (pre-RLVR) model.

  - Pass@1 : single-sample accuracy
    - 모델이 단 한번의 시도로 정답을 맞출 확률
  - Pass@K : existence of a correct path (guessing-prone)
    - 모델이 문제에 대해 K번 시도 했을 때, 정답을 맞출 확률



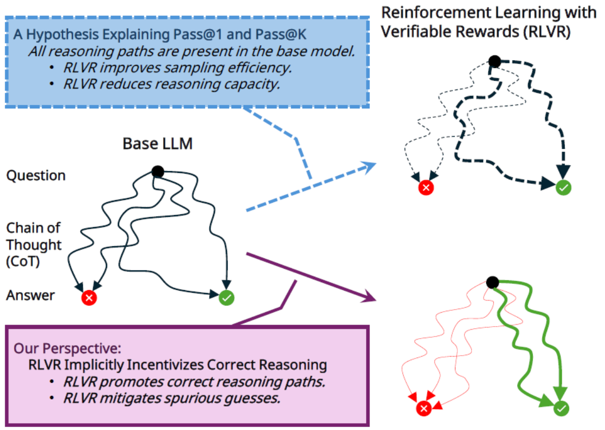
- Hypothesis (Prior work, Yue et al.)
  : All correct reasoning paths are already present in the base model, and RLVR merely improves sampling efficiency at the cost of reducing overall reasoning capacity.

→ <u>No systematic explanation exists</u>

Introduction

- Contribution

1. A systematic evaluation revealing the extended reasoning capability boundary after RLVR for both code and math tasks.
2. A theoretical understanding of why RLVR works with only answer correctness as a reward and how RLVR incentivizes correct reasoning.
3. An analysis of RLVR's training dynamics, delving deeper into optimization effects, generalization behaviors, and current limitations.
4. Confirmation of the quality improvement in reasoning CoT from learning perspective : if supervised learning on some CoT data results in better generalization on test sets, regard as high quality.



RLVR

- Since the release of DeepSeek-R1(2025), A surge of research interest in the RLVR paradigm.
- Due to the high computational cost of RLVR,
  Most studies have focused on small-, medium-sized models. (up to 32B parameters)
- However, only a few studies have addressed the theorical foundations of RLVR.
- This works emphasizes that implicitly incentivizes the correctness of reasoning paths.

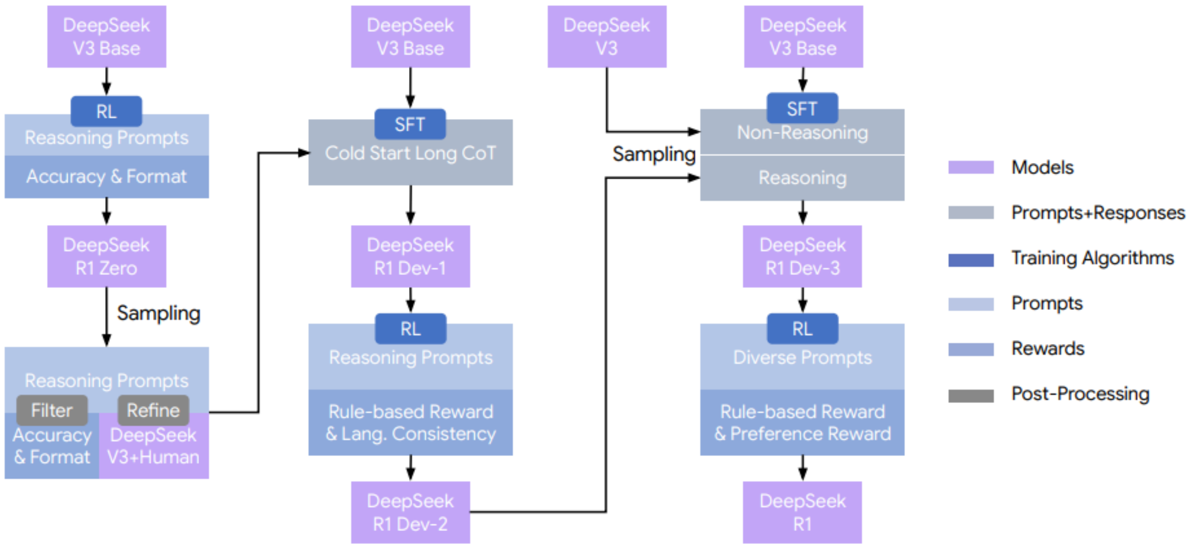


Figure 2 | The multi-stage pipeline of DeepSeek-R1. A detailed background on DeepSeek-V3 Base and DeepSeek-V3 is provided in Supplementary A.1. The models DeepSeek-R1 Dev1, Dev2, and Dev3 represent intermediate checkpoints within this pipeline.

https://arxiv.org/pdf/2501.12948

Importance of Correct CoTs

- Recent studies focus on defining synthetic reasoning tasks
  - ; Artificially constructed task where correctness of reasoning CoTs can be verified easily.
- However, it is difficult to apply to unstructured reasoning scenarios, such as math and code.
- In this work, argue that the LLM-as-a-CoT-Judge paradigm could paly a crucial role in more general reasoning tasks.

1) Extended reasoning capability boundary after RLVR

- Present **concrete benchmark evaluation** (math and code domain)
  that demonstrate how RLVR can fundamentally enhance the reasoning abilities of LLMs.

- Math
  - Correctness is judged by extracted answer token
  - High likelihood of guessing

- Code
  - Correctness is verified by actual code execution
  - Guessing is significantly reduced

Math Reasoning

- Revisiting the Pass@K Experiments conducted on the open-source model, DAPO-Qwen-32B
  - Using the based LLM, Qwen2.5-32B, curated set of 17k mathematical problems.
  - Pass@K performance of base LLMs on math reasoning can be unreliable.
    (Base LLM are capable of producing incorrect CoT yet coincidentally arriving at the ground Truth, especially under large K.)
- Introduction of a novel evaluation metric, CoT-Pass@K
  - Evaluate success only when **both the final answer and the intermediate reasoning CoT are correct**
  - Pass@1 : Single-sample accuracy
    - 모델이 단 한번의 시도로 정답을 맞출 확률
  - Pass@K : Existence of a correct path (guessing-prone)
    - 모델이 문제에 대해 K번 시도 했을 때, 정답을 맞출 확률
  - CoT-Pass@K : Correct answer + Correct reasoning

Math Reasoning

- Verifier: DeepSeek-R1-0528-Qwen3-8B
- CoT correctness strategies:
  - Any-correct: at least one verification returns correct
  - All-correct: all verifications must return correct
  - Majority-correct: the majority vote determines the outcome.
  - Manual inspection: Pass@K metrics == small positive, CoT-Pass@K metrics = 0

Math Reasoning

- Top row(Pass@K): Performance of the base LLM quickly catches up with and even surpasses the post-RLVR model as K increase.
- Bottom row(CoT-Pass@K): revealing a consistent performance gap between the models across al values of K (up to 1024)
  - AIME 2025: released after the based model's training cutoff.

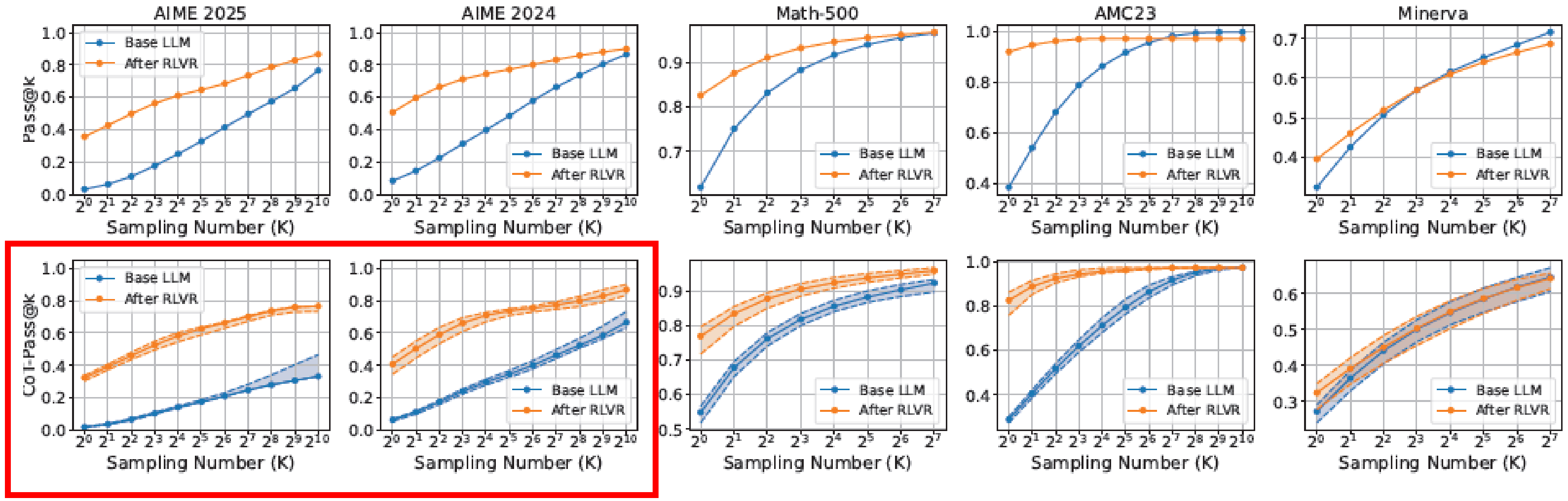


Figure 2: Comparisons of Pass@K (the top row) and CoT-Pass@K (the bottom row) on five math benchmarks (different columns) to show how RLVR could improve base LLMs. Here the base LLM is Qwen2.5-32B, and the post-RLVR model is DAPO-Qwen-32B. For CoT-Pass@K, we perform multiple verifications for each CoT using DeepSeek-R1-0528-Qwen3-8B, and display the results determined by *any-correct*, *all-correct*, and *majority-correct* strategies, which constitute the shaded area in lower subplots.

Math Reasoning

- Math-500, AMC23
  - LLM이 풀기 쉬운 문제 or 문제 일부가 pretraining data 포함
- Minerva
  - 물리 문제 중심(domain mismatch)의 bench mark

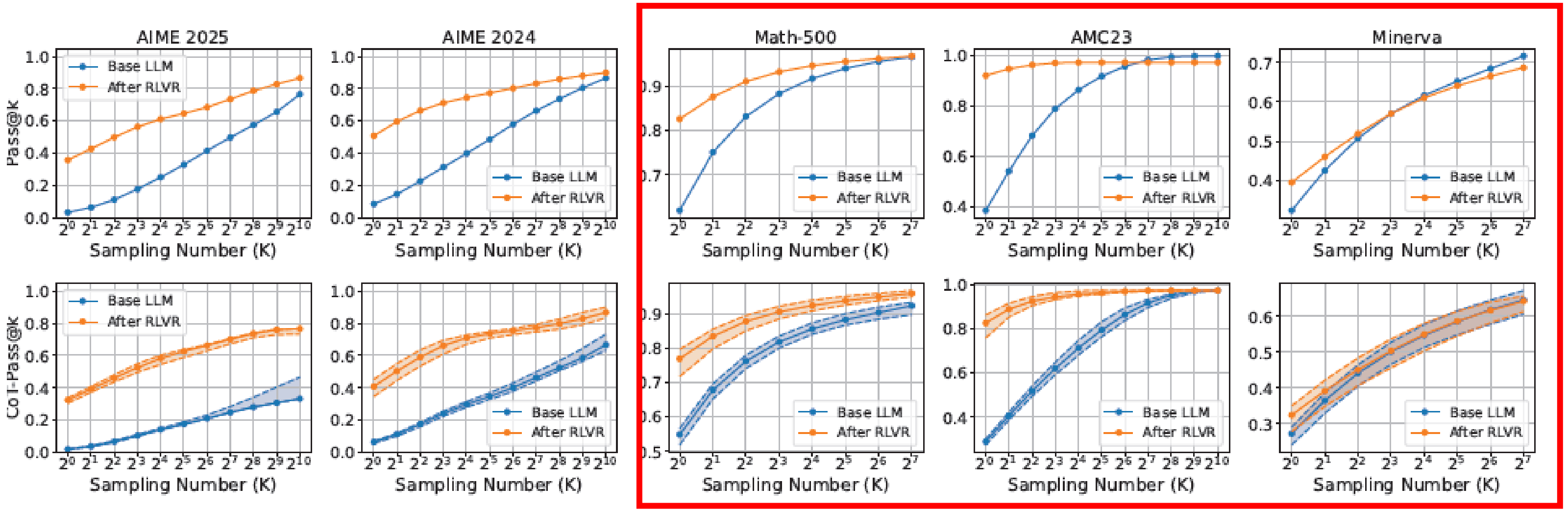


Figure 2: Comparisons of Pass@K (the top row) and CoT-Pass@K (the bottom row) on five math benchmarks (different columns) to show how RLVR could improve base LLMs. Here the base LLM is Qwen2.5-32B, and the post-RLVR model is DAPO-Qwen-32B. For CoT-Pass@K, we perform multiple verifications for each CoT using DeepSeek-R1-0528-Qwen3-8B, and display the results determined by *any-correct*, *all-correct*, and *majority-correct* strategies, which constitute the shaded area in lower subplots.

Code Reasoning

- Reproduce the Pass@K experiments across different version of LiveCodeBench
- DeepSeek-R1-Distill-Qwen-7B: pre-RLVR distillation model, already strong at reasoning
- AceReason-Nemotron-7B: RLVR → extend the reasoning capability boundary

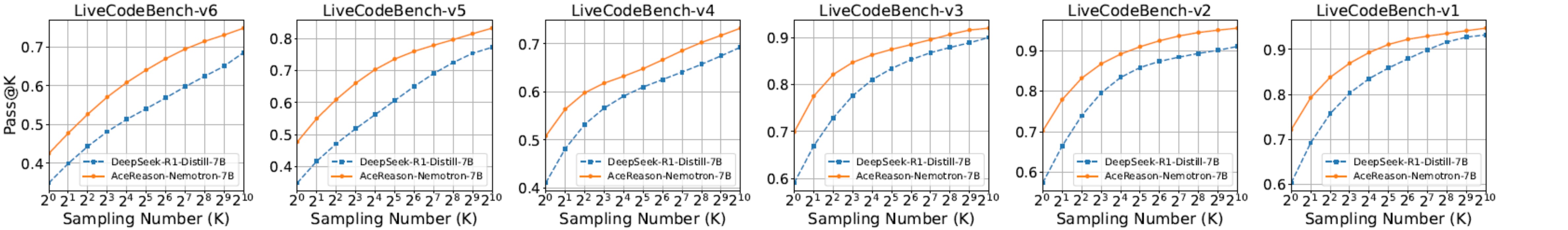


Figure 3: Comparisons of Pass@K across six LiveCodeBench versions to show how much RLVR could enhance distilled LLMs. Here the distilled LLM is DeepSeek-R1-Distill-Qwen-7B, and the post-RLVR model is AceReason-Nemotron-7B.

2) Theoretical Understanding of RLVR for LLMs

- How RLVR fundamentally incentivizes correct reasoning for pre-trained language models?
- Problem setup
  - Question prompt q, sample G responses Y = $\{y_1, y_2, y_3, \ldots, y_G\}$ from policy π_θ
  - π_θ is an LLM model parameterized by θ
  - c_i be the CoT in response y_i, and a_i be the final answer

$$\mathcal{I}_{\text{CoT}}(c_i) = \begin{cases} 1 & \text{if } c_i \text{ is correct} \\ 0 & \text{otherwise} \end{cases}, \quad \mathcal{I}_{\text{Ans}}(a_i) = \begin{cases} 1 & \text{if } a_i \text{ is correct} \\ 0 & \text{otherwise} \end{cases}.$$

  - The CoT correctness $I_{CoT}(c_i)$: the intermediate tokens of a response(c_i) expressing necessary and accurate logics that lead to the ground truth.
  - $P_c^\theta = P_{\pi_\theta}(I_{CoT}(c) = 1)$: probability of generating a correct CoT
  - **The answer correctness $\boldsymbol{I_{Ans}(a_i)}$:** be verified programmatically.
  - **Verifiable reward $\boldsymbol{R(y_i)}$** is binary and determined solely by answer correctness : $R(y_i) = I_{Ans}(a_i)$

2) Theoretical Understanding of RLVR for LLMs

- Problem setup
  - GRPO advantage

Verifiable reward $R(y_i)$

$$\hat{A}(y_i) = \frac{R(y_i) - \mu_{\mathbf{Y}}}{\sigma_{\mathbf{Y}}}, \quad \mu_{\mathbf{Y}} = \frac{1}{G}\sum_{j=1}^{G} R(y_j), \quad \sigma_{\mathbf{Y}} = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(R(y_j) - \mu_{\mathbf{Y}})^2}.$$

$$\nabla_\theta J(\theta) \approx \frac{1}{G}\sum_{i=1}^{G} \hat{A}(y_i)\nabla_\theta \log \pi_\theta(y_i \mid q).$$

2) Theoretical Understanding of RLVR for LLMs

- Assumptions
  - pre-trained LLMs have established strong knowledge and logic priors.
  - Decoupling CoT and answer correctness, Introduce a critical ***Logic prior*** assumption:
  - Compared with incorrect CoTs, correct CoTs have high probabilities to induce correct answers.

$$P(\mathcal{I}_{\text{Ans}}(a_i) = 1 \mid \mathcal{I}_{\text{CoT}}(c_i) = 1) = \alpha > P(\mathcal{I}_{\text{Ans}}(a_i) = 1 \mid \mathcal{I}_{\text{CoT}}(c_i) = 0) = \beta.$$

- Theorem (GRPO Implicitly Incentivizes Correct Reasoning)
  - the GPRO increase the probability of generating correct CoTs(P_c^θ) in next round

$$\mathbb{E}\left[\hat{A}(y_i) \mid \mathcal{I}_{CoT}(c_i) = 1\right] > 0, \quad \mathbb{E}\left[\hat{A}(y_i) \mid \mathcal{I}_{CoT}(c_i) = 0\right] < 0,$$

3) Training Dynamics of RLVR

- Analyze RLVR training dynamics by reproduced DAPO training.
- key indicators
  - for each prompt q sampled with G responses,
  - **the number of answer passes**:

$$C = \sum_{i=1}^{G} \mathcal{I}_{\text{Ans}}(a_i) \qquad P(CA)^{(q)} = \frac{C}{G}$$

  - **the number of both CoT and answer passes**:

$$D = \sum_{i=1}^{G} \mathcal{I}_{\text{CoT}}(c_i) \cdot \mathcal{I}_{\text{Ans}}(a_i) \qquad P(CC|CA)^{(q)} = \frac{D}{C}$$

3) Training Dynamics of RLVR

- Optimization Effects
  - The probability of generating correct answers for these quest almost reach 1.
  - Producing more correct reasoning CoT increase
  - RLVR not only optimizes the final verifiable reward but also implicitly incentivizes correct reasoning.

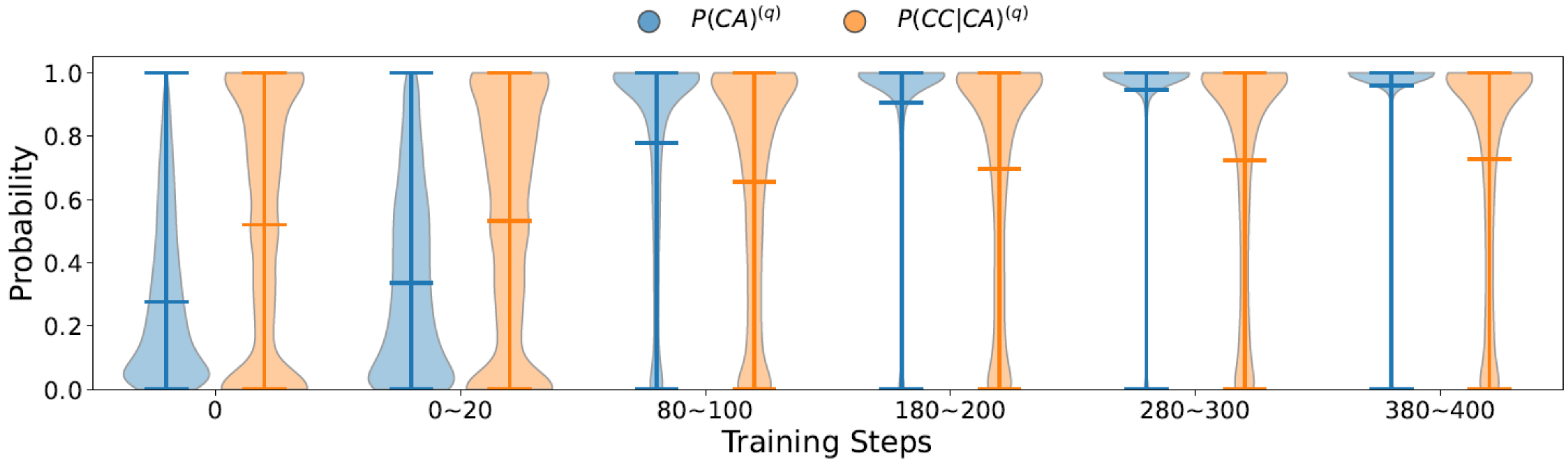


Figure 4: The evolution of $P(CA)^{(q)}$ (the fraction of correct answers for prompt q) and $P(CC|CA)^{(q)}$ (the fraction of correct CoTs within the correct answers for prompt q) for fully optimized training questions over the course of DAPO training.

3) Training Dynamics of RLVR

- Generalization Behaviors
  - Leads to generalization improvement of both Pass@K and CoT-Pass@K <u>from very beginning.</u>

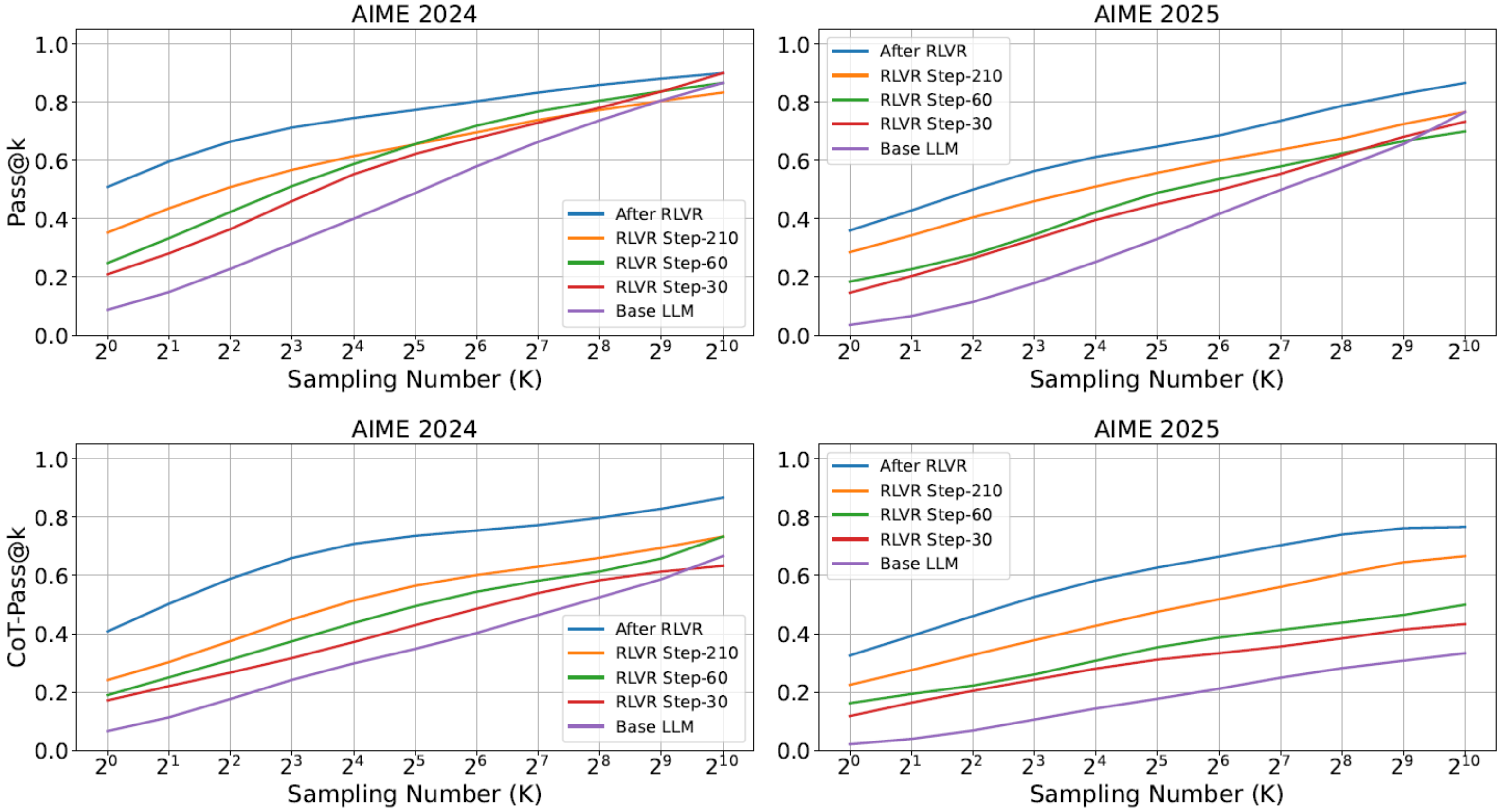


Figure 5: The evolution of Pass@K (the top row) and CoT-Pass@K (the bottom row) performance on AIME 2024 and 2025 for different model checkpoints during the DAPO training.

3) Training Dynamics of RLVR

- Limitation of DAPO
  - $P(CA)^{(q)}$ approaches 1.0 after 400 steps → no longer learnable using GRPO advantage
  - $P(CC|CA)^{(q)}$ is around 0.7 → still observe non-negligible portion of imperfect CoTs

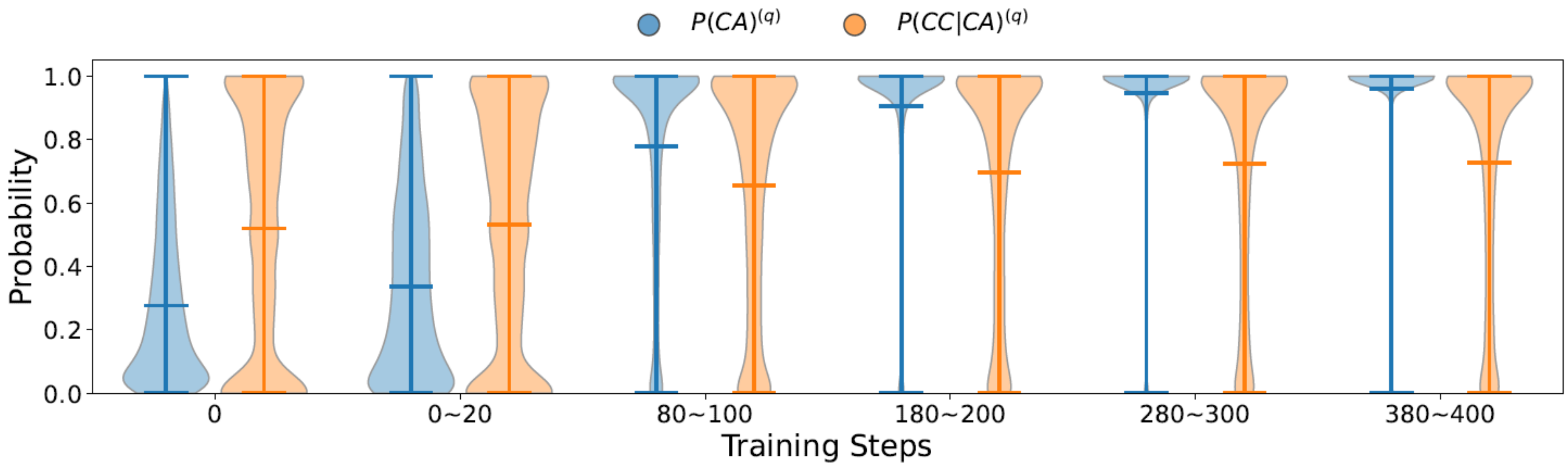
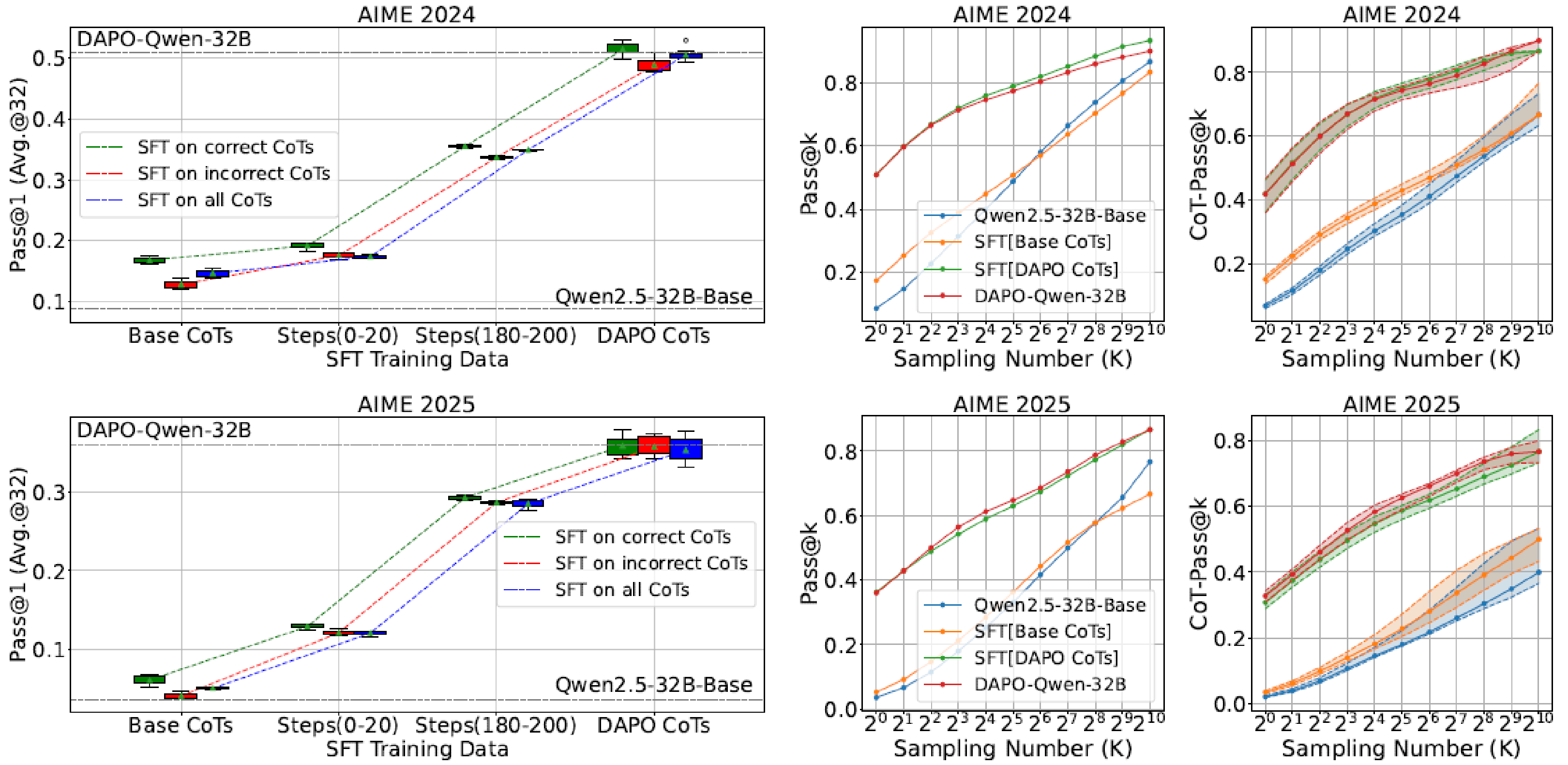


Figure 4: The evolution of $P(CA)^{(q)}$ (the fraction of correct answers for prompt q) and $P(CC|CA)^{(q)}$ (the fraction of correct CoTs within the correct answers for prompt q) for fully optimized training questions over the course of DAPO training.

4) The Quality of Reasoning CoTs Enhance By RLVR

- Leverage supervised fine-tuning(SFT) to assess the quality of reasoning CoTs enhanced by RLVR.
- If the CoT data is of high quality, expect the post-SFT model to exhibit improved generalization performance.



(a) The CoT quality at different RLVR stages, using Pass@1 on test sets as the proxy metric.

(b) The CoT quality before and after RLVR, using (CoT-)Pass@K on test sets as the proxy metric.

Conclusion

- Addresses whether RLVR genuinely incentivizes novel reasoning in base LLMs.
    - implicit incentivization of correct reasoning (theory)
    - early generalization during training (dynamics)
    - high-quality CoTs reusable via SFT (quality)