Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?

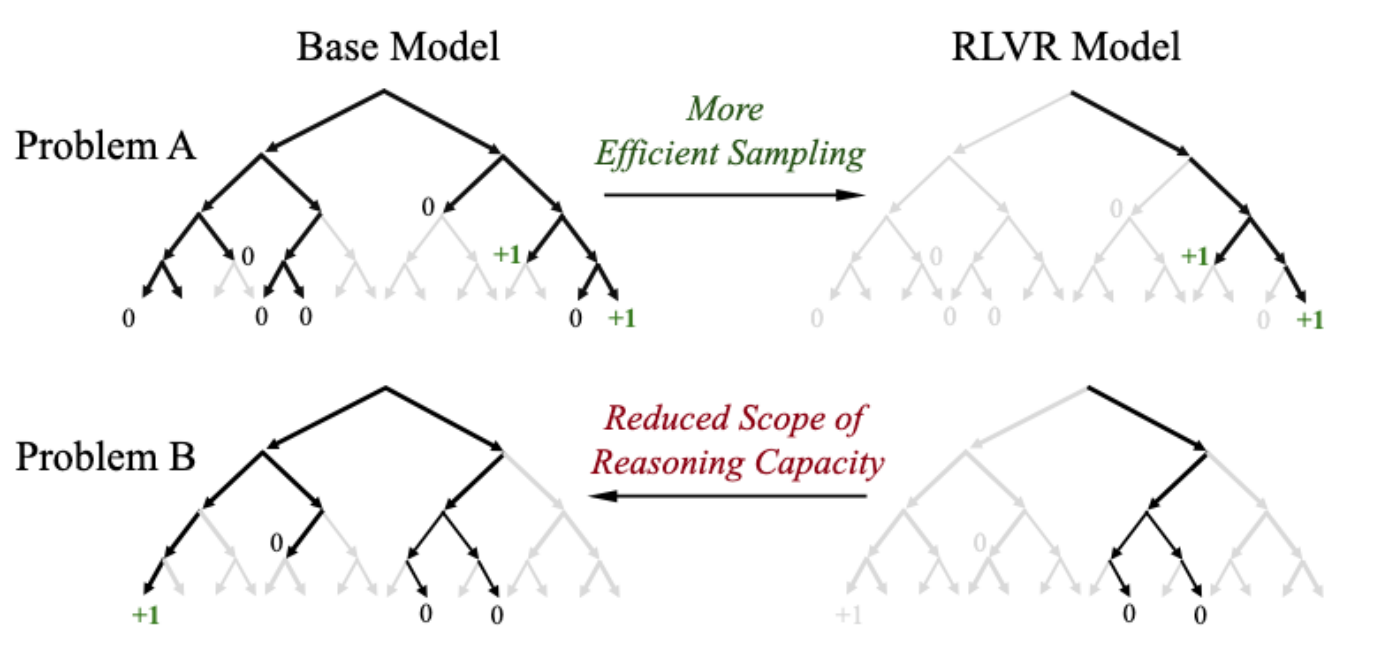
Yang Yue et al., NeurIPS 2025

Kevin Liu Pratham Hemlani

March 11, 2025

EECE 571 — University of British Columbia

Problem Setup and Central Question



Common interpretation of RLVR

- Better pass@1 is often read as evidence that RLVR has taught the model new reasoning ability.
- The usual intuition is that reinforcement learning should discover strategies beyond the starting policy.

Question asked by this paper

- Does RLVR actually expand the model's **reasoning capacity boundary**?
- Or does it mainly make **already-existing correct paths** easier to sample from the base model?

Preliminaries: RLVR, Zero-RL Training, and the Comparison Setting

RLVR in this paper

- RLVR trains the current model using **verifiable rewards**: mathematical correctness, compilation, or unit-test success.
- The paper focuses on policy-gradient-style methods trained from **on-policy samples**.

Zero-RL setting

- For mathematics, the paper follows the **zero-RL** setting: RL is applied directly to the **pretrained base model**, without SFT.
- For coding and visual reasoning, prior open-source work usually starts from **instruction-tuned** models because pure zero-RL is less stable.

Comparison principle

- The key comparison is always between a **base model** and its **RLVR-trained counterpart** under the same evaluation setup.

Preliminaries: Why the Paper Uses pass@k

Definition for a single problem

- Sample k outputs from the model.
- The question counts as solved if **at least one** sampled output is correct under the verifier.
- Dataset pass@k is then the proportion of problems solvable within k trials.

Why pass@k matters here

- Greedy or average sampled accuracy only reflect **average-case behavior**.
- This paper wants to probe the model's **reasoning capacity boundary**: given enough attempts, what problems can the model solve at all?
- Large-k pass@k is therefore used as a **coverage metric**.

Interpretation

- If RLVR truly creates new reasoning ability, it should remain stronger even when the base model is allowed many samples.

Preliminaries: How pass@k Is Estimated in Practice

Why estimation matters

- Directly re-estimating pass@k separately for every value of k would be noisy and expensive.
- The paper therefore uses the **unbiased low-variance estimator**.

What they do

- For each question, they sample n responses once, where n is the **largest** value of k used in the curve.
- They count how many of those n responses are correct.
- From that one sample set, they estimate pass@k for **all smaller** $k \leq n$.

$$\text{pass@}k := \mathbb{E}_{x_i \sim \mathcal{D}} \left[1 - \frac{\binom{n-c_i}{k}}{\binom{n}{k}} \right]$$

Comparison with Best-of-N and Majority Voting

Metric	What it measures	Weakness/Strength
pass@1	One-shot accuracy	Reflects average-case behavior, but can miss rare successful traces.
Majority vote	Selection quality after sampling	May fail to select a correct solution even when that solution appears.
pass@k	Coverage of potentially solvable problems	Best evaluate whether RLVR expands the set of reachable correct solutions.

What the authors emphasize

- Best-of-N and majority voting are useful for **practical answer selection**.
- But this paper is asking a different question: **capacity coverage**, not deployment-time selection.

Random Guessing Issue and Manual CoT Validation

Why this is a real concern

- In mathematics, a model may output an **incorrect chain of thought** but still get correct by chance, especially when k becomes large.
- So a raw pass@k number is only persuasive if those large-k successes correspond to **valid reasoning traces**, not lucky answer hacks.

How the paper addresses it

- For coding, strong compilers and unit tests make accidental success much less plausible.
- For math, the authors manually inspect CoTs on the **hardest low-accuracy questions**.
- On GSM8K, the base model solves 25 such questions and 24 contain at least one correct CoT; the RL model also solves 25, and 23 contain at least one correct CoT.
- On filtered AIME24 hard cases, the base model solves 7 such questions, with 5 of 6 non-ambiguous cases containing at least one correct CoT; the RL model solves 6, with 4 containing at least one correct CoT.

Experimental Setup: Models, Benchmarks, and Domains

Task	Start Model	RL Framework	RL Algorithm(s)	Benchmark(s)
Mathematics	LLaMA-3.1-8B Qwen2.5-7B/14B/32B-Base Qwen2.5-Math-7B	SimpleRLZoo Oat-Zero DAPO	GRPO	GSM8K, MATH500 Minerva, Olympiad AIME24, AMC23
Code Generation	Qwen2.5-7B-Instruct DeepSeek-R1-Distill-Qwen-14B	Code-R1 DeepCoder	GRPO	LiveCodeBench HumanEval+
Visual Reasoning	Qwen2.5-VL-7B	EasyR1	GRPO	MathVista MathVision
Deep Analysis	Qwen2.5-7B-Base Qwen2.5-7B-Instruct DeepSeek-R1-Distill-Qwen-7B	VeRL	PPO, GRPO Reinforce++ RLOO, ReMax, DAPO	Omni-Math-Rule MATH500

Scope of evaluation

- **Domains:** mathematics, code generation, visual reasoning, and deep analysis
- **Model families:** Qwen, LLaMA, DeepSeek-based distilled models, and Qwen-VL
- **Training pipelines:** SimpleRLZoo, Oat-Zero, DAPO, Code-R1, EasyR1, and VeRL

Why this setup matters

- The paper is not making a one-model or one-benchmark claim.
- The central pattern is tested across **multiple tasks, model sizes, and RL pipelines**.

Experimental Protocol and Evaluation Fairness

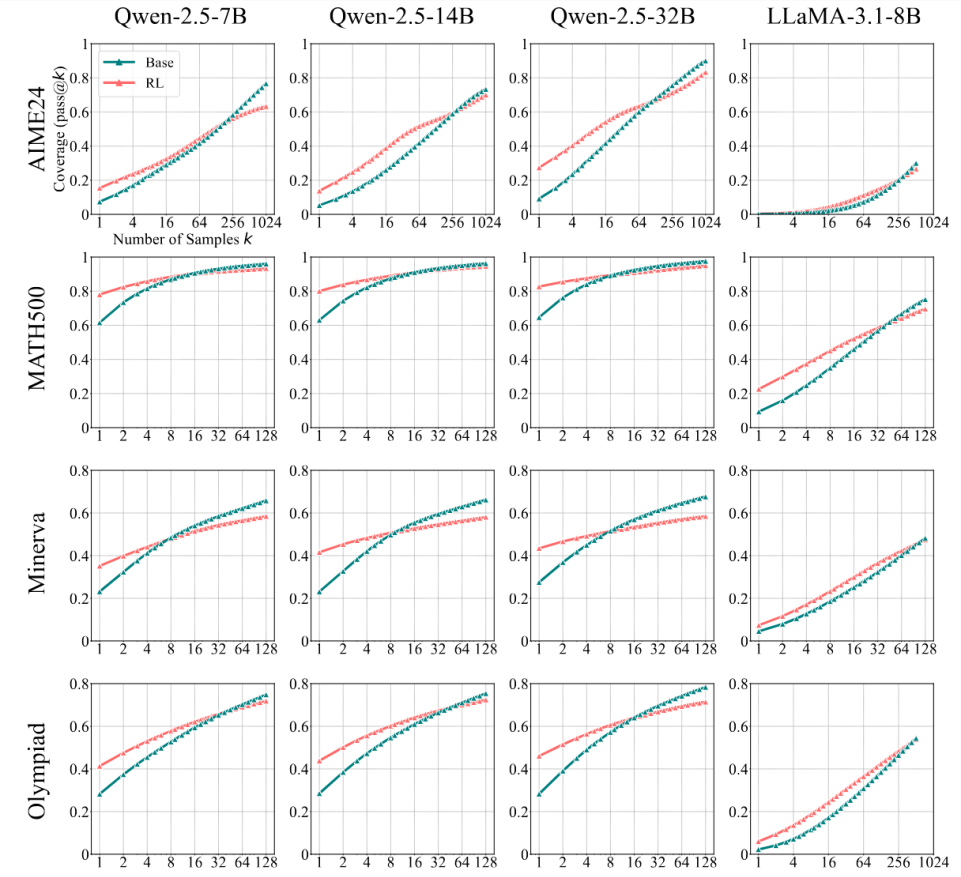
Protocol details highlighted by the paper

- Temperature $= 0.6$, top-$p = 0.95$
- Maximum generation length up to 16,384 tokens
- Base models are evaluated **without few-shot prompting**
- Base and RLVR models use the **same zero-shot prompt** as in RL training, or the benchmark's default prompt

Why this matters

- It removes the confound that few-shot examples might artificially enlarge the base model's reachable set.
- The comparison is therefore designed to isolate the effect of **RLVR itself**, rather than prompt engineering.

Current RLVR Models Often Exhibit Narrower Reasoning Coverage Than Their Base Models



- At small k, RLVR models usually outperform their base models, which matches the standard impression that RLVR helps.
- As k increases, base models repeatedly **catch up and often surpass** their RLVR counterparts.

The Crossover Also Appears in Code Generation and Visual Reasoning

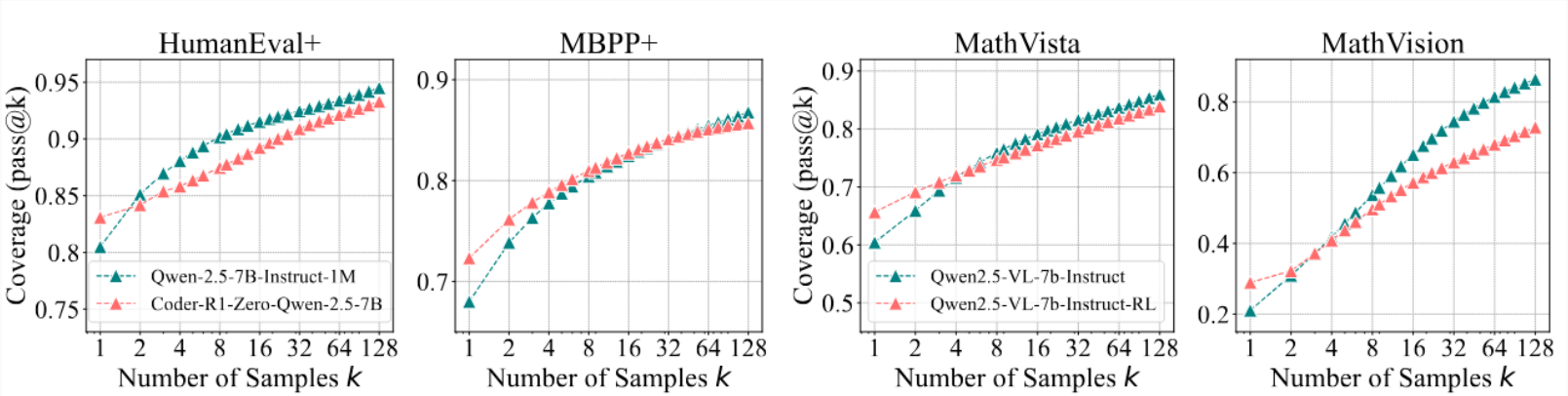
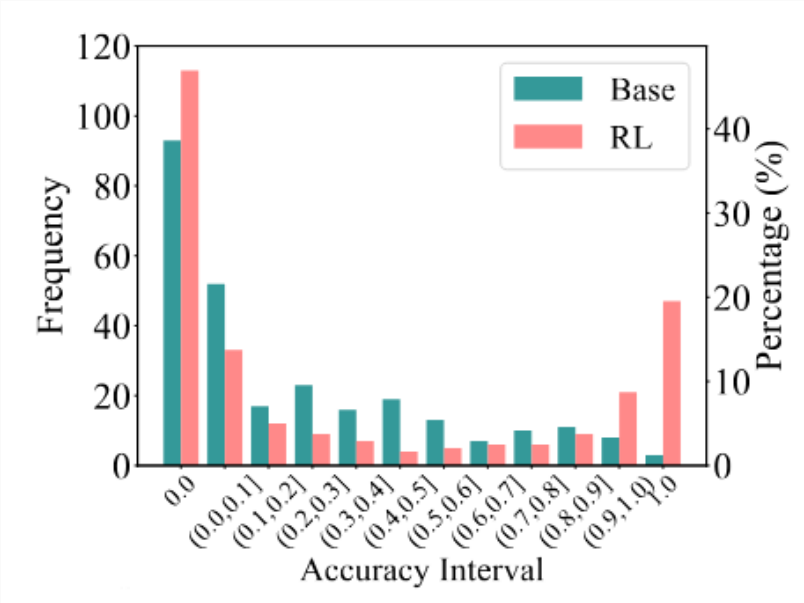


Figure 4: Pass@k curves of base and RLVR models. **(Left)** Code Generation. **(Right)** Visual Reasoning.

- Figure 4 shows that the same qualitative pattern is not confined to mathematics.
- In coding, the use of compilers and unit tests makes accidental success less plausible.
- So the paper's central claim is not merely a math-specific artifact; it appears across multiple reasoning domains.

Reasoning Paths Already Present in Base Models: Accuracy Histogram



What Figure 5 shows

- After RLVR, more mass shifts toward **high-accuracy bins**.
- But the mass at accuracy $= 0$ also **increases**.

Interpretation

- Some previously solvable problems effectively drop out of coverage.
- At the same time, RLVR improves sampling efficiency on problems that were already within reach.
- This is why average accuracy can improve even while reasoning coverage narrows.

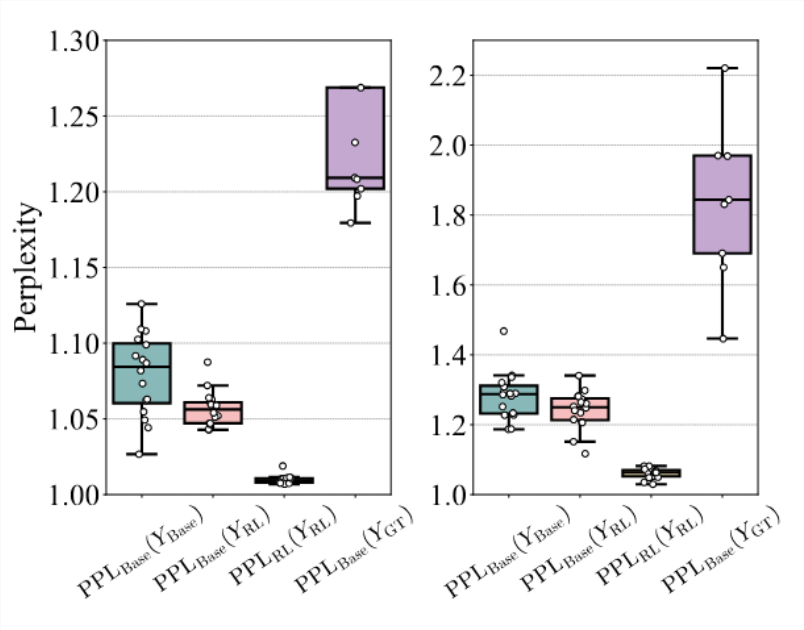
Reasoning Paths Already Present in Base Models: Solvable-Problem Coverage (Table 2)

Base	SimpleRLZoo	AIME24	MATH500
✓	✓	63.3%	92.4%
✓	✗	13.3%	3.6%
✗	✓	0.0%	1.0%
✗	✗	23.3%	3.0%

- On **AIME24**, the fraction solved by RLVR but not by the base model is reported as **0.0%**.
- On **MATH500**, the corresponding RL-only fraction is only **1.0%**.
- By contrast, there remains a non-trivial **Base-only** portion.

Interpretation. The cleanest reading is that the RLVR-solvable set is much closer to a **subset of the base-model-solvable set** than to an expansion.

Reasoning Paths Already Present in Base Models: Perplexity Analysis



What Figure 6 tests

- Perplexity represents the model's **predictive ability** for this text sequence, and the low·er the model can understand the text better.

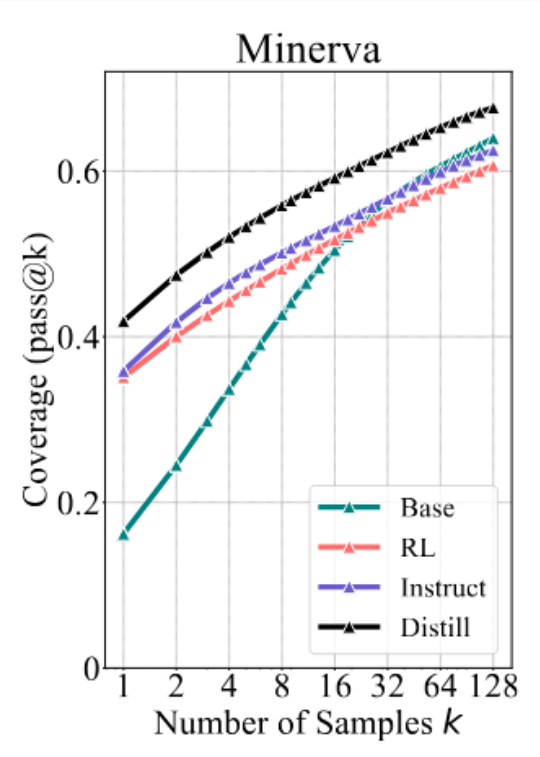
Observed pattern

- $\mathrm{PPL}_{\mathrm{Base}}(Y_{\mathrm{RL}} \mid x)$ stays close to the **low-perplexity region** of base model responses.

Interpretation

- RLVR is **sharpening the base prior** rather than moving beyond it.

Distillation Expands the Reasoning Boundary



Why Figure 7 matters

- The distilled model stays **consistently above** base, instruct, and RL across pass@k.
- This is qualitatively different from the RLVR crossover pattern.

Interpretation

- Distillation can import **new reasoning patterns** from a stronger teacher.
- In the paper's framing, distillation can genuinely **expand the reasoning boundary**, while current RLVR remains bounded by the base model.

Measuring Sampling Efficiency: The Δ_{SE} Metric

How far is RLVR from fully exploiting the base model?

Sampling Efficiency Gap:

$$\Delta_{SE} = \text{Base pass@256} - \text{RLVR pass@1}$$

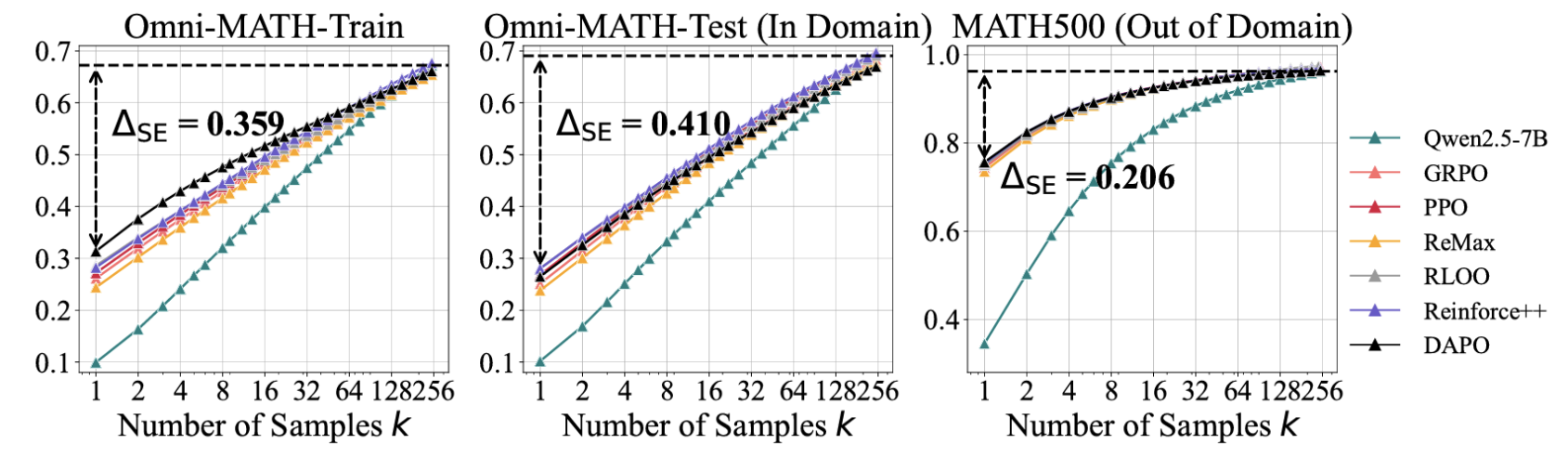
Lower is better — zero means perfect exploitation.

Tested across 6 algorithms: PPO, GRPO, Reinforce++, RLOO, ReMax, DAPO

Sampling Efficiency Gap Persists Across All Algorithms

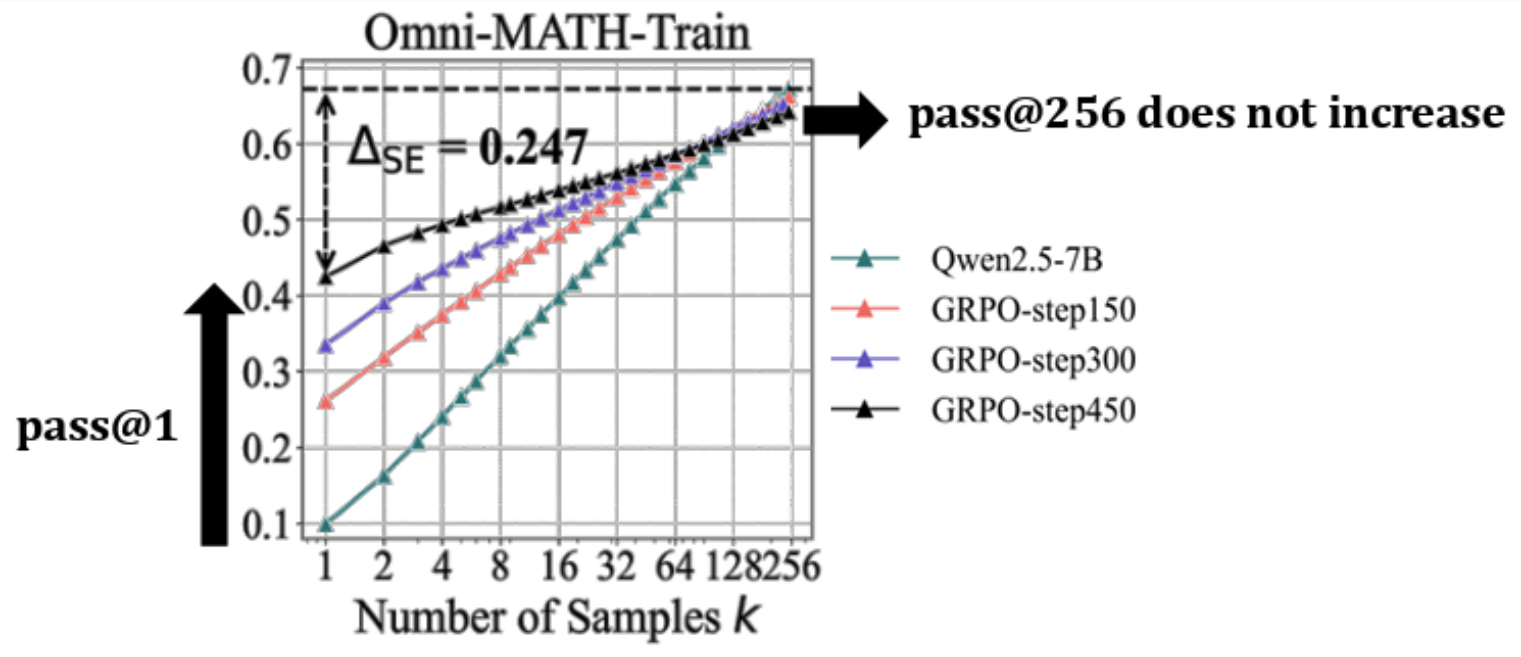
- Δ_{SE} ranges 42.6 to 43.9 on in-domain test set
- Despite maximum freedom to deviate (KL removed), all algorithms cluster at the same Δ_{SE}

Novel algorithms alone are insufficient to close the gap



More Training Does Not Close the Gap

- pass@1 consistently improves as training progresses (steps 150 to 450)
- pass@256 does not increase — the ceiling stays fixed
- More training specializes the model without expanding what it can solve



Ruling Out Alternative Explanations

Explanation	Test	Verdict
KL Regularization	Add KL penalty	× **Makes it worse**
Reduced Entropy	Raise temperature	× **Partial only**
Model Scale	Frontier model	× **Same pattern**

Why is RLVR Bounded by the Base Model?

	Traditional RL	RLVR for LLMs
Action Space	Small, structured	Exponentially larger
Starting Point	From scratch	Pretrained prior
Upper Bound	None	Bounded by prior

The prior is a double-edged sword:

- Enables exploration in a vast action space. Without it, training from scratch is intractable.
- Constrains reasoning. Deviating from the prior produces incoherent outputs and negative reward, pushing the policy back toward what the base model already knows.

Strengths

- **Broad and consistent empirical scope.** Multiple model families, six algorithms, three reasoning domains — crossover pattern holds across all variations.
- **Pass@k as an evaluation lens for reasoning coverage.** Standard accuracy metrics would have missed this — large k separates sampling efficiency from reasoning coverage.
- **Distillation as a positive control.** If you only show RLVR fails, a skeptic says your evaluation is broken. Shows a method that genuinely expands coverage, ruling out the possibility the evaluation itself is flawed.

Weaknesses: Empirical

- **Primary experiments on 7–32B models.** Training data is also narrow (GSM8K+MATH, LeetCode+TACO) — at frontier scale with richer data, properties of models can change suddenly
- **Rebuttal adds 70B+ results, but thinly.** Tulu-3-70B evaluated on only 100 MATH500 problems — not enough to draw strong conclusions about whether the pattern fully holds at that scale.
- **Scale experiment is preliminary.** Magistral-Medium model size undisclosed, not reasoning SOTA — authors acknowledge this themselves.

Weaknesses: Theoretical

- **Restricted to on-policy policy gradient methods.** All six algorithms tested share the same fundamental mechanism. Off-policy or value-based RL approaches could offer different learning dynamics and may not face the same exploration bottleneck.
- **Future directions remain speculative.** The paper identifies the problem clearly but proposed solutions are unproven — none have been shown to reliably overcome the limitation in practice.

What We Learned

- **Evaluation design shapes conclusions.** pass@1 and pass@k at large k tell fundamentally different stories — reflects assumptions about what capability means.
- **The pretrained prior is both the enabler and the bottleneck.** This framing applies broadly — any fine-tuning method that relies on the model's own outputs faces the same structural tension, prior shapes what the model can discover.
- **Negative results require strong experimental design.** Distillation as a positive control is what makes this paper credible — without it the evaluation is just a negative claim with no positive baseline.
- **RLVR improves reliability, not capability.** A useful reframe — RLVR makes the model more consistently correct, not fundamentally smarter.

Open Questions

- **Does the pattern hold at full frontier scale?** Magistral-Medium is a proxy — testing on models like DeepSeek-R1-Zero directly requires impractical compute.
- **Would process rewards change the conclusion?** Outcome-based rewards are binary and sparse. Step-level rewards could in principle guide exploration beyond the prior — but current process reward models are noisy and hackable.
- **Is the reasoning boundary the right metric?** One reviewer raised this directly — pass@k at large k may not reflect practical deployment value given inference costs.
- **What does this mean for models trained with mixed objectives?** Qwen3-235B combines RLVR and long-context CoT SFT — disentangling their contributions is currently infeasible.

Suggestions

- **Strengthen the scale experiments.** 70B rebuttal results are promising but thin — a full pass@k evaluation on larger models with sufficient samples would significantly strengthen the generalizability claim.
- **Report correct solutions with flawed CoTs.** Manual CoT validation covers a small sample. A larger-scale analysis would tighten the random guessing bound and strengthen the claim that pass@k measures genuine reasoning coverage.
- **Explore process reward ablation.** All experiments use binary outcome-based rewards. Swapping these for step-level process rewards under the same pass@k setup would test whether the conclusion holds beyond this specific reward structure.

Future Directions

- **Process rewards and credit assignment.** Step-level reward signals could guide exploration without relying entirely on the prior — requires better reward models that are robust to hacking.
- **Agentic and multi-turn RL.** Allow the model to interact with external tools, retrieve information, and receive environment feedback across turns — expands the reasoning space beyond single-pass generation.
- **Curriculum and meta-learning.** Build generalizable reasoning skills progressively rather than optimizing directly for final answers — may help the model escape the prior's gravity over time.