Reinforcement Learning with Verifiable Rewards

Madhav Donepudi

Tanishk Govil

David Jagga

Background

RLVR: RL where reward is **automatically checkable**

Why people care: offers a scalable way to push LLMs towards **correct** outcomes without expensive human grading

Core Pipeline: Prompt → LLM output → Verifier

Three papers:

1. **R1:** RLVR yields strong reasoning gains
2. **Boundary Paper:** asks whether gains reflect efficiency vs new capacity
3. **ProRL:** shows conditions where longer RL can expand boundaries

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI, 2025
Guo et al., arXiv:2501.12948

The Bottleneck of Supervised Fine-Tuning

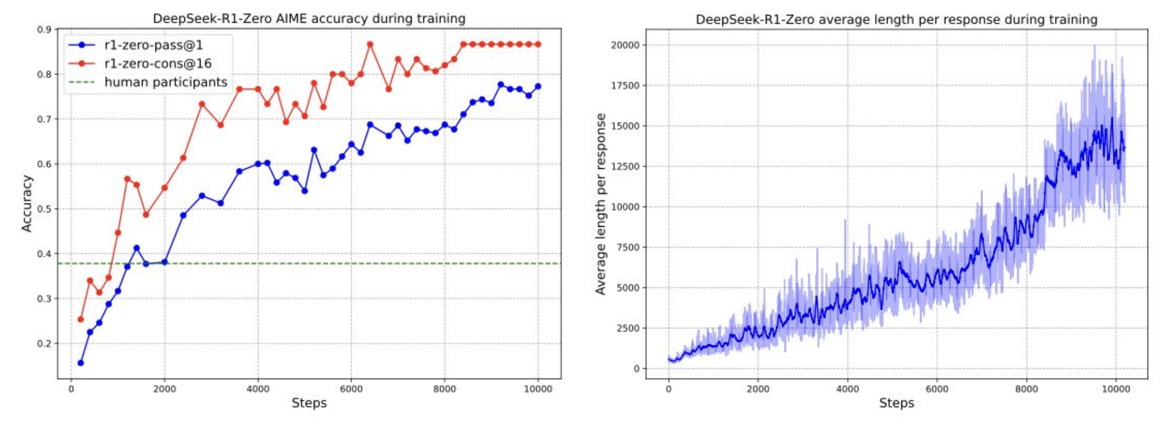
- Problem: SFT relies on expensive human data and caps performance at human levels

- Research Question: Can reasoning emerge purely from Reinforcement Learning without human-labeled trajectories?

- Hypothesis: Scaling “test-time compute” via RL allows models to discover superior, non-human reasoning paths

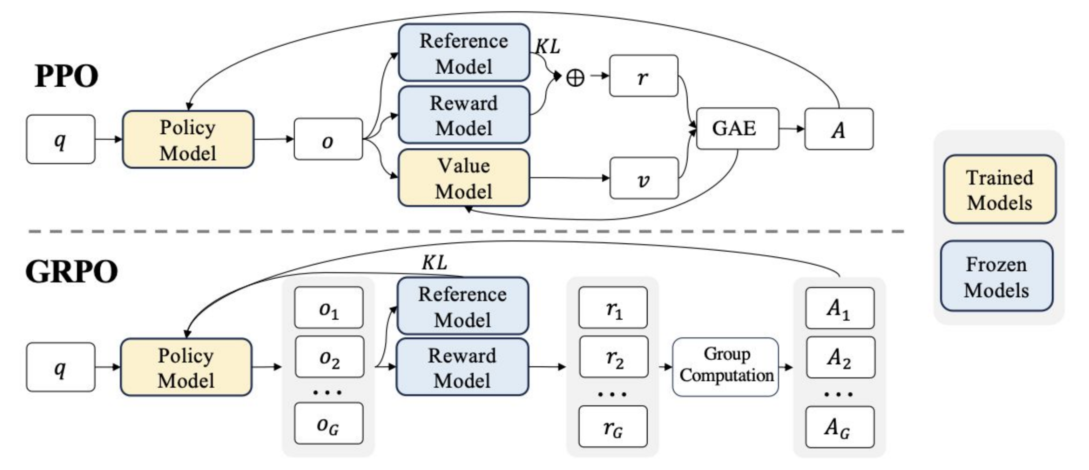
DeepSeek-R1-Zero: Pure RL Discovery

- Initialization: Trained directly from DeepSeek-V3-Base with Zero SFT
- Reward Structure: Solely rule based (Accuracy + Format)
- Accuracy:
  - Math: Correct answer presented in the correct format (eg. within a box)
  - Code: Compiler verifies if code passes specific test cases
- Format:
  - Structural Check: Encourages thinking process within mandatory <think> and <answer> tags
- Optimization: Pure exploration and no human style guidance



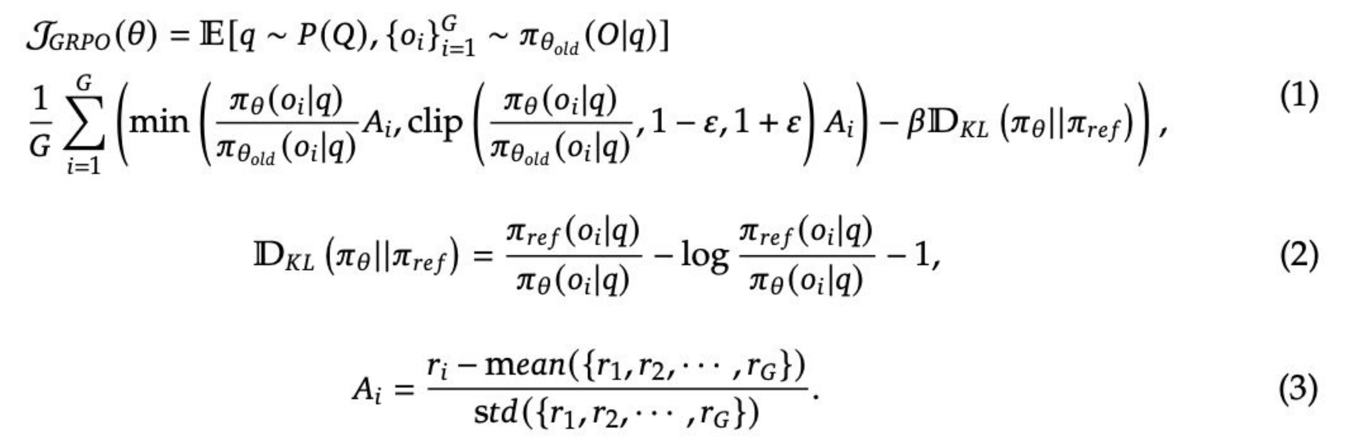
Group Relative Policy Optimization (GRPO)

- Motivation: Standard PPO is too memory intensive due to the “Value Model”
- Innovation: Advantages are computed relative to a group of outputs
- Impact: Significant VRAM savings which enables large scale training of long-chain reasoning



Mathematical Framework of GRPO

- Advantage: Standardized reward within a group of G outputs using advantage normalization
  - Correct response: Non-negative Advantage
  - Incorrect response: Non-positive Advantage
- KL Penalty: Unbiased estimator used to maintain policy stability



$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right), \tag{1}$$

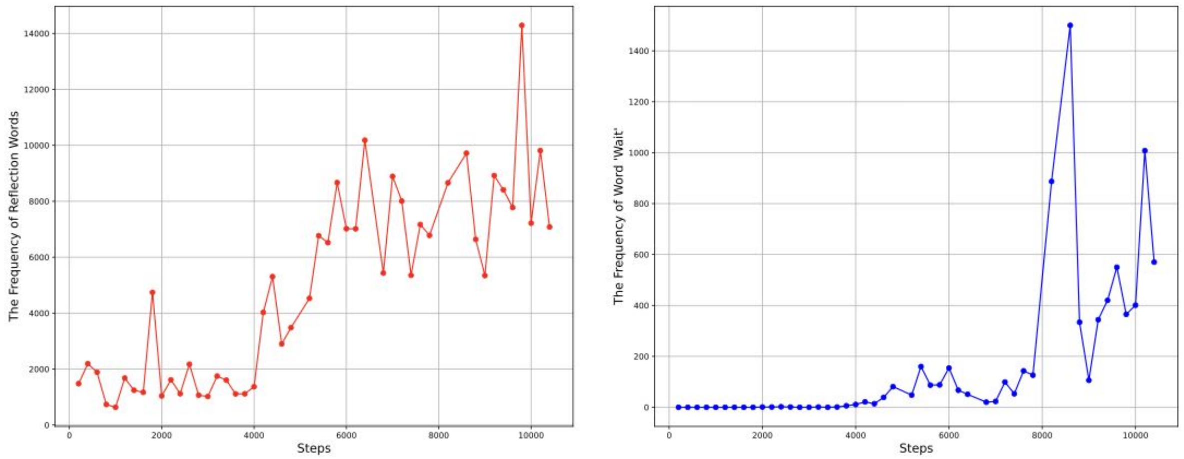
$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \tag{2}$$

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}. \tag{3}$$

1. GRPO Objective
2. KL Penalty Term
3. Advantage from verifier rewards

Emergent Self-Correction and Scaling

- Discovery: Model autonomously learned to backtrack, verify, and re-evaluate
- Linguistic Markers: Sharp increase in “reflective” tokens (e.g., “Wait”, “Let me re-check”)



The “Aha Moment”

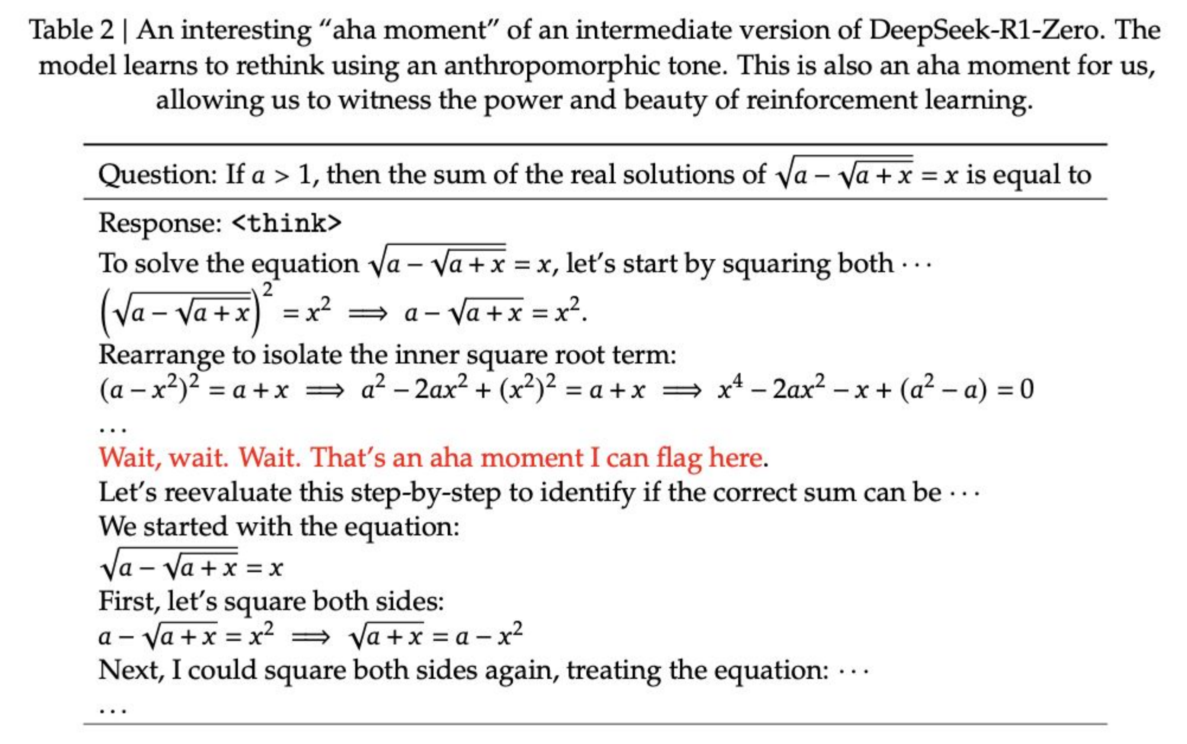


Table 2 | An interesting “aha moment” of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

---

Response: `<think>`

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

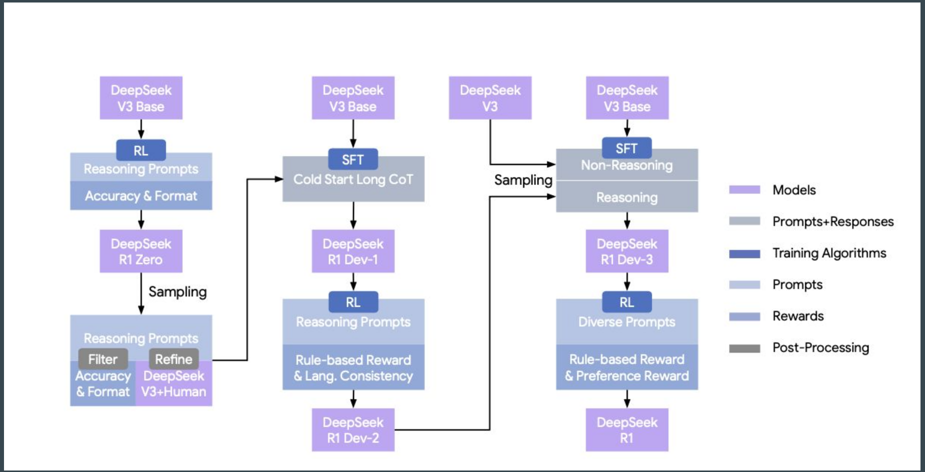
$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

...

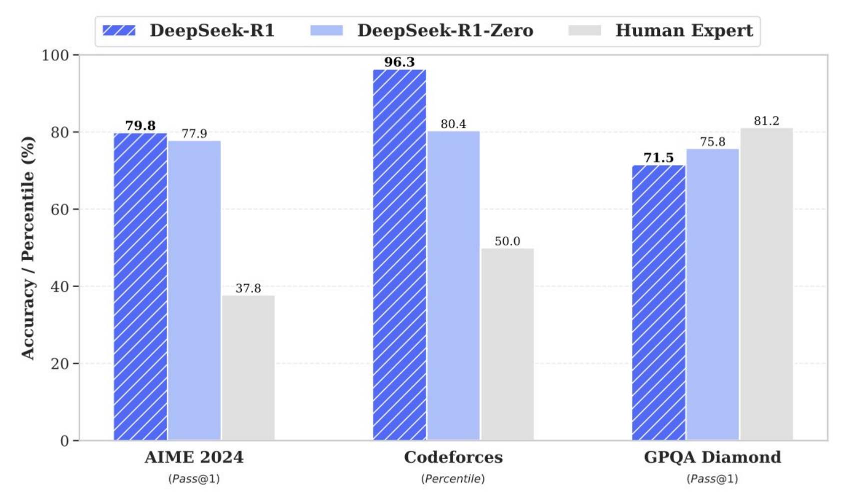
The DeepSeek-R1 Training Pipeline

- Stage 1 (Cold Start): SFT on ~5k long-COT samples for readability
- Stage 2 (Reasoning RL): Large scale RL focused on STEM + Logic
- Stage 3 (Rejection Sampling): Synthetic data generation (800k samples)
- Stage 4 (Preference RL): Alignment for helpfulness and safety



Benchmarking Reasoning Capabilities

- Competitive Math: 79.8% on AIME 2024
- Competitive Coding: 96.3% human percentile on Codeforces
- Conclusion: Achieves similar performance with OpenAI-o1-1217 using significantly less compute



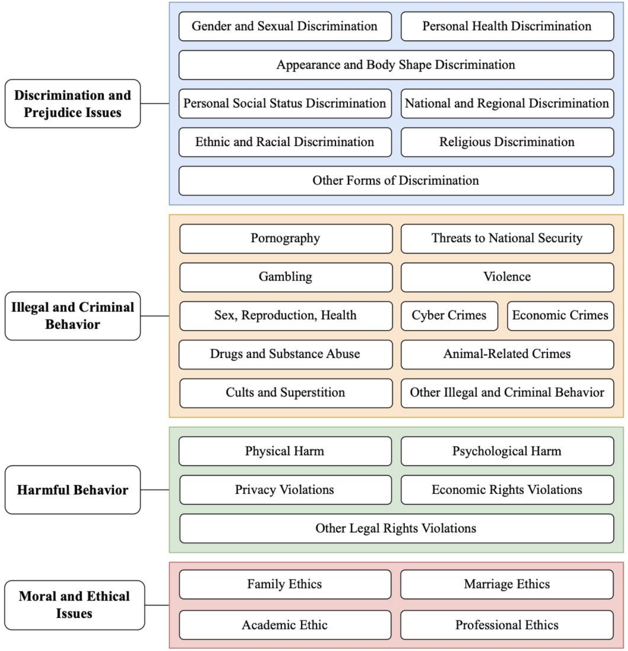
Knowledge Distillation

- Data Gathering: Used DeepSeek-R1 to generate ~800k high-quality “reasoning traces”
- Filtering: Only “correct” answers with readable logic were kept
- Fine-Tuning: These traces were used for Supervised Fine-Tuning on smaller models
- Efficiency: A distilled 1.5B model outperforms non-reasoning GPT-4o on mathematical benchmarks

Safety Evaluation and Risk Control

- Point: The model implements a “Safety Reward” during the final RL stage to mitigate risks without compromising reasoning

- System: Use of an external risk review system (sending queries to DeepSeek-V3) to detect and retract unsafe dialogue



- **Discrimination and Prejudice Issues**
  - Gender and Sexual Discrimination
  - Personal Health Discrimination
  - Appearance and Body Shape Discrimination
  - Personal Social Status Discrimination
  - National and Regional Discrimination
  - Ethnic and Racial Discrimination
  - Religious Discrimination
  - Other Forms of Discrimination
- **Illegal and Criminal Behavior**
  - Pornography
  - Threats to National Security
  - Gambling
  - Violence
  - Sex, Reproduction, Health
  - Cyber Crimes
  - Economic Crimes
  - Drugs and Substance Abuse
  - Animal-Related Crimes
  - Cults and Superstition
  - Other Illegal and Criminal Behavior
- **Harmful Behavior**
  - Physical Harm
  - Psychological Harm
  - Privacy Violations
  - Economic Rights Violations
  - Other Legal Rights Violations
- **Moral and Ethical Issues**
  - Family Ethics
  - Marriage Ethics
  - Academic Ethic
  - Professional Ethics

Figure 13 | Taxonomy of in-house safety benchmark.

Unsuccessful Approaches

- Process Reward Models (PRM): Deemed inefficient due to the difficulty of defining universal “correct steps” and the high risk of reward hacking

- Monte Carlo Tree Search (MCTS): Challenges in scaling due to the exponentially larger search space of token generation

Limitations

- Inference Efficiency: Models lack an internal “stop signal,” often overthinking trivial tasks and wasting compute
- Domain Constraints: Current success is tied to “verifiable” tasks (STEM) and extending this to subjective fields remains a challenge
- Linguistic Consistency: Occasional language mixing and drops in consistency occur during long, multi-turn interactions
- Prompt Sensitivity: Performance can be inconsistent based on phrasing, and the model often requires explicit instructions to “think” for optimal results

Future work

- Adaptive Compute: Scaling reasoning time based on task difficulty to improve efficiency
- Agentic Tasks: Expanding from logic puzzles to real world engineering and autonomous planning
- Multi-Modal Reasoning: Integrating reasoning chains with vision and audio inputs for cross-modal tasks
- Nuanced Rewards: Developing reward systems for subjective tasks that lack binary answers

Conclusion

- Emergent Reasoning: Proves logic (backtracking, self-correction) can be discovered via RL without human data
- Efficiency: GRPO eliminates the Value Model, drastically reducing the hardware cost of training
- Accessible AI: Distillation successfully transfers traces to small, efficient models
- A New Scaling Rule: Intelligence now scales with test-time compute rather than just model size.

DeepSeek-R1 provides a blueprint for next-gen AI, replacing human-led imitation with autonomous, machine native logic

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

Yue et al.
arXiv:2504.13837

Overview

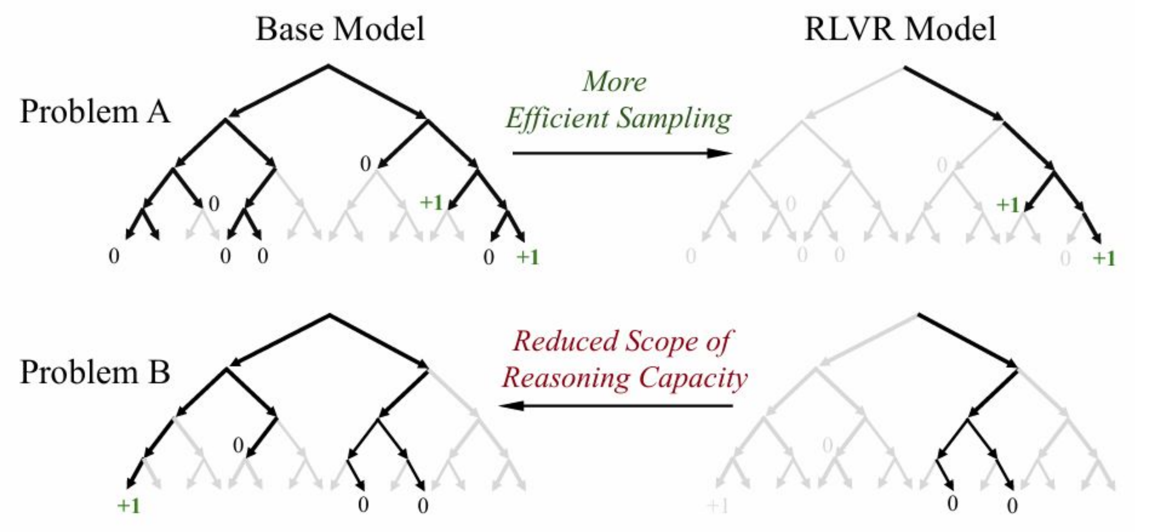
- Common belief: RL helps agents discover new strategies
  - Go
  - Atari

- People expect RL with verifiable rewards (RLVR) to “self-improve” like classic RL

Main Research Question: Does current RLVR discover *new* reasoning patterns, or just exploit base-model ones?

Core Intuition: Efficiency vs Boundary

- Sampling efficiency: how likely the model reaches a correct path in few tries
- Reasoning boundary/coverage: what fraction of problems are solvable by the model given many tries



Key Metric: pass@k

pass@k = “can we solve it if we try enough times?”

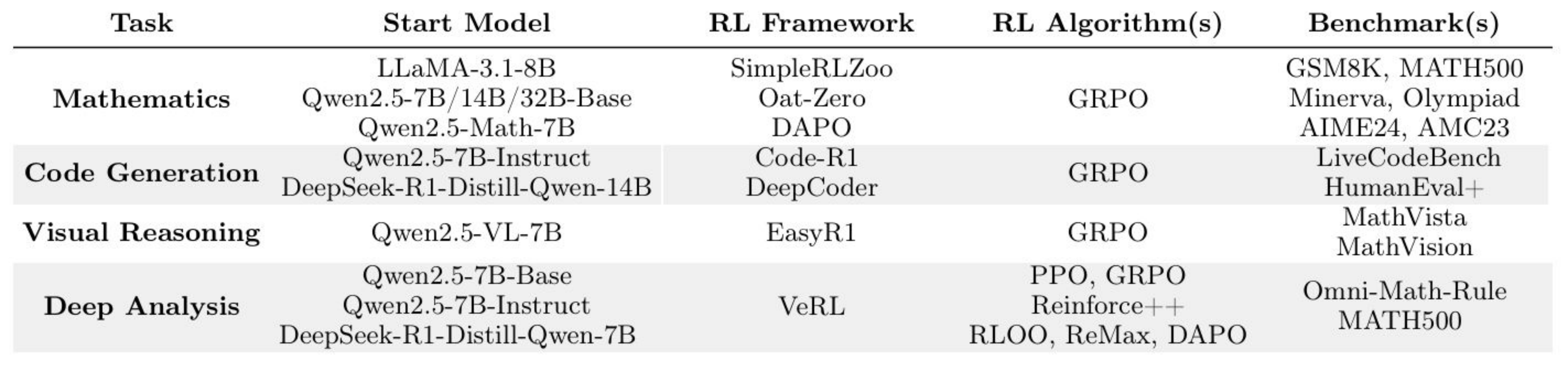
- Sample k outputs; pass@k = 1 if any of them are correct
- Estimated by sampling k completions using temperature-based sampling

Purpose:

- Evaluating boundaries
- If RL truly adds reasoning ability, it should solve problems the base model can’t
- **Efficiency** = better pass@1
- **Boundary** = better pass@k as k → large

Experiments at a Glance

- Test across multiple model families + RL algorithms
- Goal: find general patterns across several types

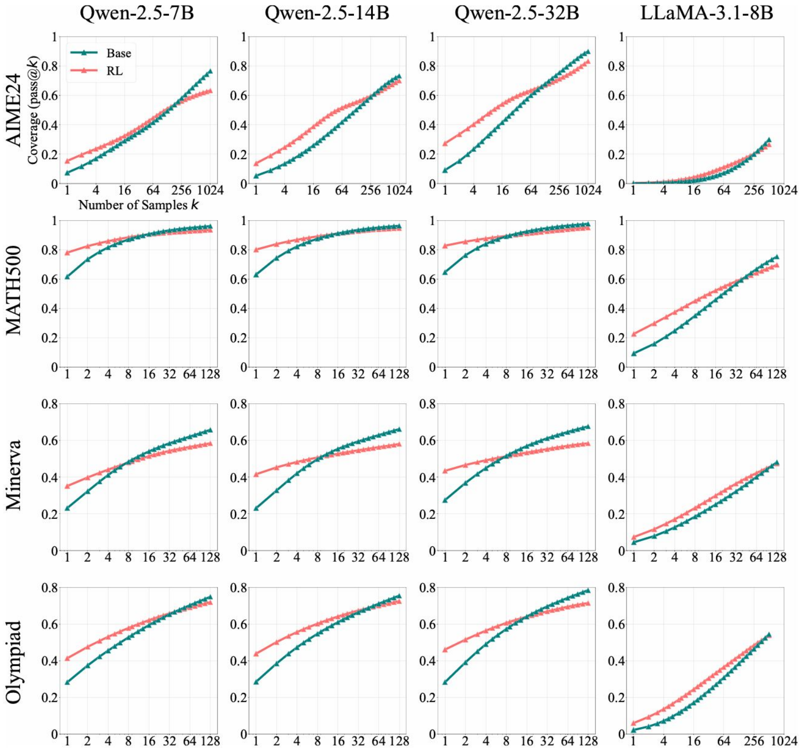


Task	Start Model	RL Framework	RL Algorithm(s)	Benchmark(s)
Mathematics	LLaMA-3.1-8B Qwen2.5-7B/14B/32B-Base Qwen2.5-Math-7B	SimpleRLZoo Oat-Zero DAPO	GRPO	GSM8K, MATH500 Minerva, Olympiad AIME24, AMC23
Code Generation	Qwen2.5-7B-Instruct DeepSeek-R1-Distill-Qwen-14B	Code-R1 DeepCoder	GRPO	LiveCodeBench 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

Main result 1: Math pass@k curves flip

- RLVR > base at small k (pass@1)
- Base > RLVR at larger k

Bottom Line: RL helps get answers faster, but the base model often has *broader coverage* when you sample enough



Main result 2: Same pattern in code + visual

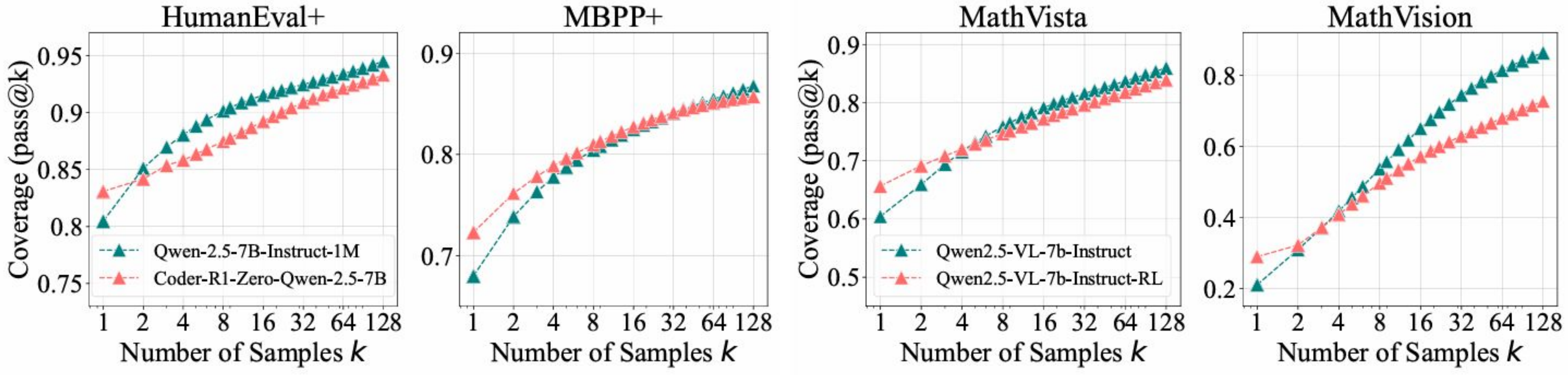


Figure 4: Pass@k curves of base and RLVR models. **(Left)** Code Generation. **(Right)** Visual Reasoning.

Same behavior found outside math. Base model surpasses RLVR at higher k, suggesting RLVR isn't expanding set of solvable problems

Then why do accuracy averages increase?

Observations:

- More mass near accuracy ~1.0
- Fewer low accuracies
- BUT more accuracy = 0

Interpretation:

- RLVR makes some problems unsolvable
- Polarizes probabilities; hyper-focuses on already good strategies

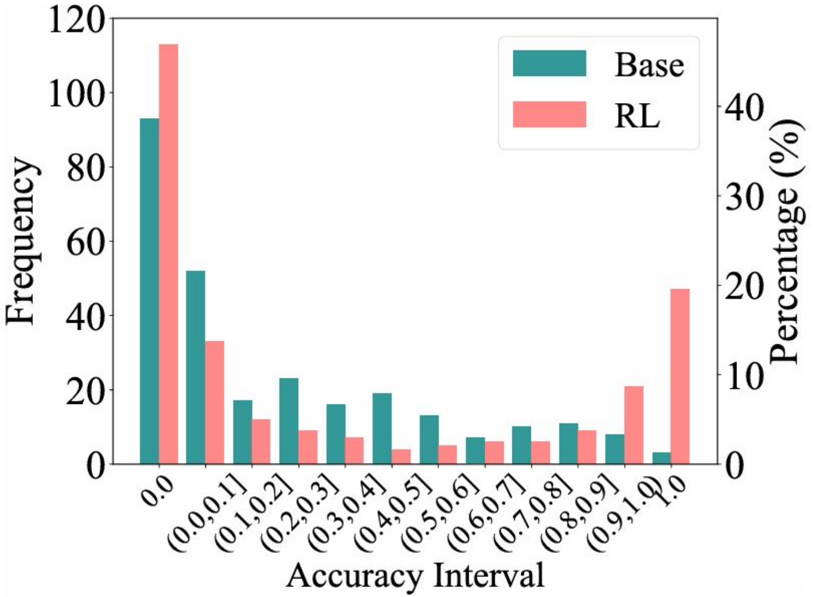
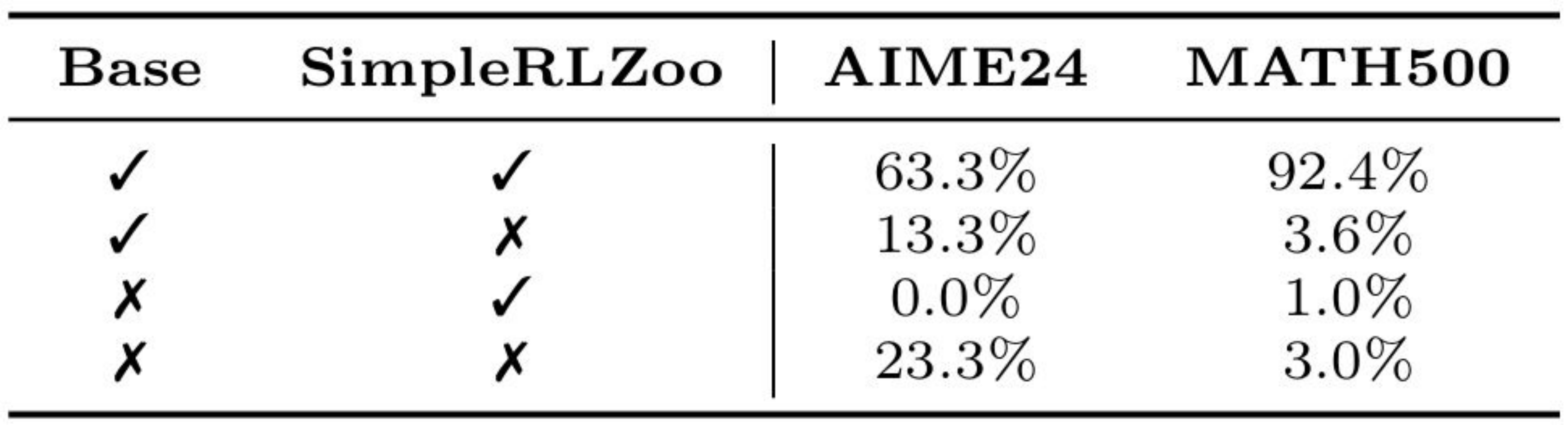


Figure 5: Qwen2.5-7B Accuracy Histogram on Minerva.

Coverage analysis: RLVR solves (almost) a subset of base model



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

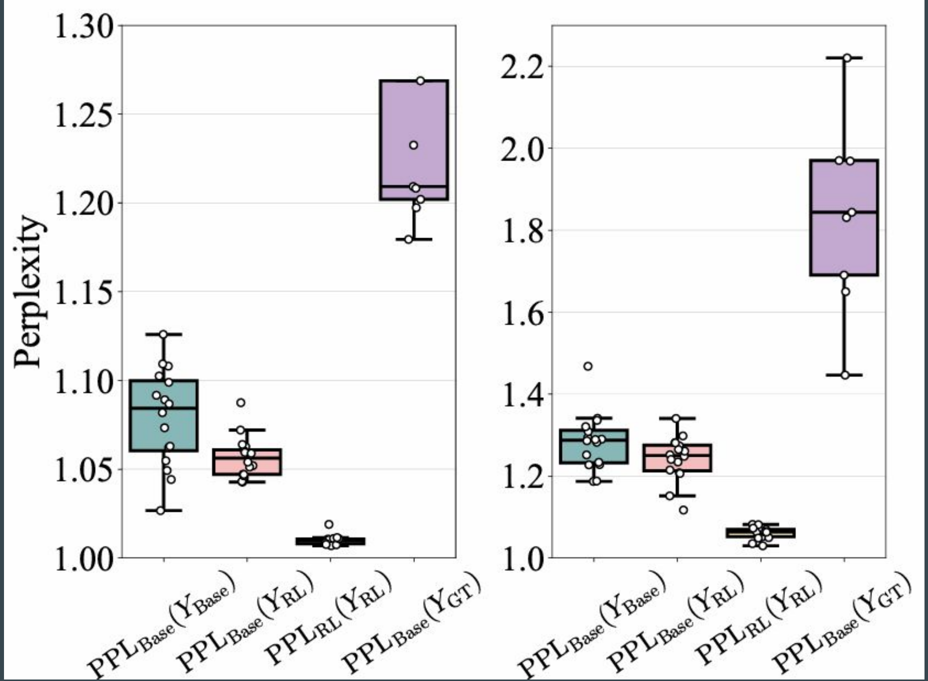
Very few cases where RLVR solves something the base model can’t, but many the other way around.

Perplexity test: Are RL solutions already in the base model?

Perplexity: if base assigns low perplexity to RL-generated solutions, they are likely to be generated by the base

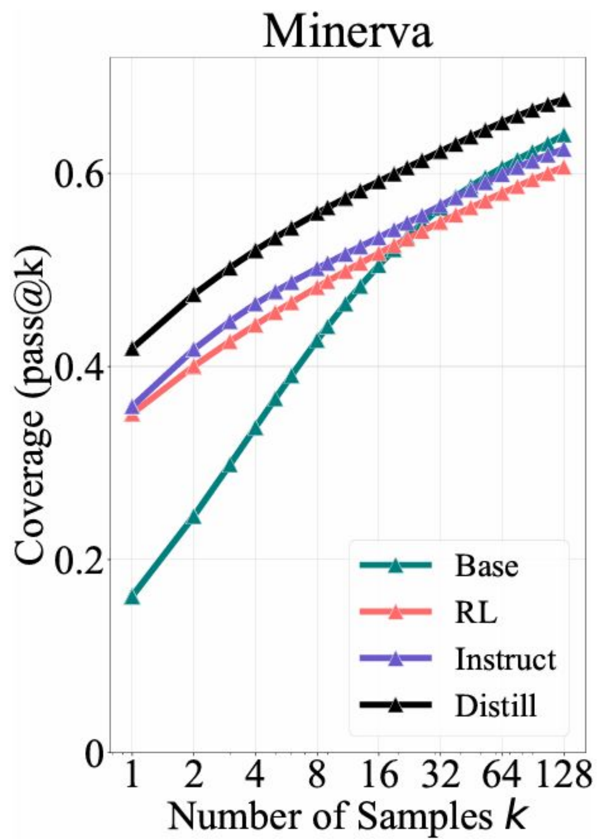
Observation: Base model is not surprised by RL's answers, so RL is exploiting existing reasoning paths

$$\mathrm{PPL}_m(\mathbf{Y} \mid x) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T} \log P(y_t \mid x, y_1, \ldots, y_{t-1})\right)$$



Distillation is different: it *can* expand the boundary

- Distilled model pass@k consistently above base model
- Imports new reasoning patterns from a stronger teacher
- Truly expands the student's reasoning boundary



Discussion: why RLVR gets stuck + how to fix it

1. Vast action space
    a. Exponentially larger than other environments
    b. Effective exploration is much harder
2. Starting from a pre trained prior:
    a. Can't start from scratch like other RL environments
    b. Struggles to explore new reasoning patterns
    c. Samples that deviate almost always get negative rewards

Future Directions:

- Fine-grained credit assignment – add intermediate signals, not just 0/1
- Scaling RL iteration – start by training on easier subproblems

Conclusion and Limitations

Main Ideas:

1. RLVR boosts pass@1 efficiency
2. Base can surpass at large k → boundary not expanded
3. RL reasoning paths often already present in base model
4. Re-weights distribution making some reasoning paths extremely unlikely

Limitations:

1. Many top models and training pipelines are proprietary
2. Field is moving fast, emerging techniques may solve this

ProRL: Prolonged Reinforcement Learning Expands Reasoning Boundaries in Large Language Models

Liu et al.
arXiv:2505.24864

Background

- ProRL Methodology: Is there still any merit to RLVR training?
- Updated Training Scheme
- What was Accomplished: Can models learn something new through the use of RLVR?

Methodology: Mitigating Entropy Collapse

- **Decoupled Clipping**
    - Separates the Upper and Lower Clipping Bounds (Token Probability)
    - What does this accomplish?
- **Dynamic Sampling Policy Optimization**
    - Filtering Questions Based on Performance
- **Reference Policy Reset**
    - KL Vanishing Problem
    - KL Penalty
    - Resetting the “Anchor Point”
- **Prolonged Training**

$$\text{clip}(r_\theta(\tau), 1-\epsilon_{low}, 1+\epsilon_{high}).$$

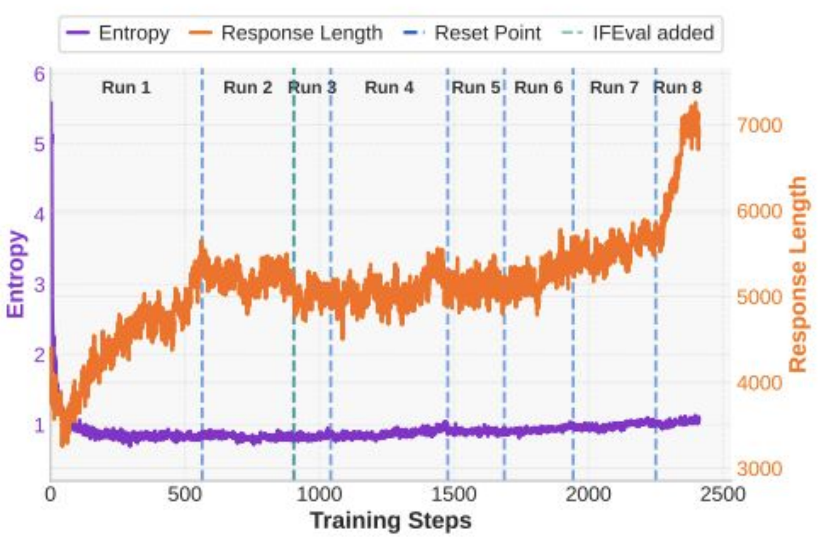


Figure 2: ProRL training dynamics.

$$L_{KL-RL}(\theta) = L_{GRPO}(\theta) - \beta D_{KL}(\pi_\theta || \pi_{ref}).$$

Task Diversity

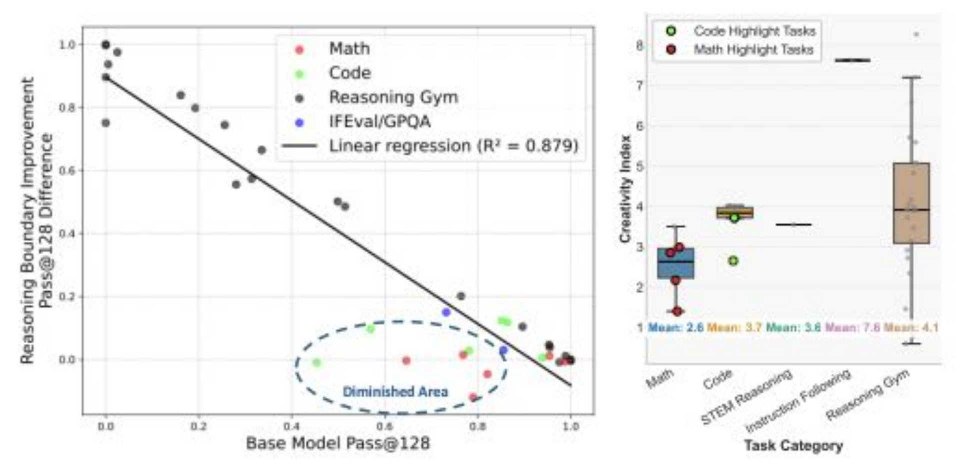
- **Math (AIME/AMC):** Provides strict, binary rewards to teach precision.
- **Code (Codeforces)** : Provides continuous rewards (passing test cases) to teach debugging and edge-case handling.
- **Logic (Reasoning Gym)** : Targets pure reasoning rules (e.g., Zebra puzzles) where base models are weakest.
- **STEM**: Tests Knowledge Application. Combines Specific Scientific Knowledge with reasoning for better real-world problem representation
- **Instruction Following:** Can the model continue to follow basic instruction

Experimental Setup

- Baseline & Competitors:
    - **The Base: DeepSeek-R1-Distill-Qwen-1.5B** (The starting student model).
    - **The Specialists: DeepScaleR (Math-only) and DeepCoder (Code-only)** of the same size. ProRL beats them both (+4.6% Math, +6.5% Code).
    - **The Giant: DeepSeek-R1-Distill-Qwen-7B (5x larger)**. ProRL matches or beats this larger model despite having fewer parameters.

Key Results

- Reversing Collapse
  - Widening Distribution of “Correct Answers”
- Weak-to-Strong Generalization
  - Where does RL work best? Where is it less effective?
  - Zero-to-One: OOD capability achieved
- Creativity Index
  - DOLMA: pre training corpus
  - N-gram Overlap
  - Where was their little Creativity?



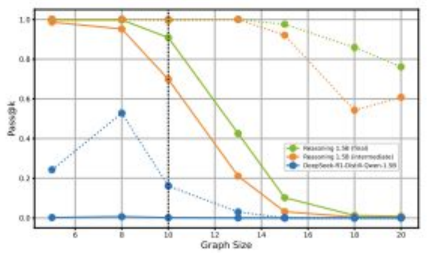


Figure 6: ProRL generalizes to increased task difficulty on *graph_color*.

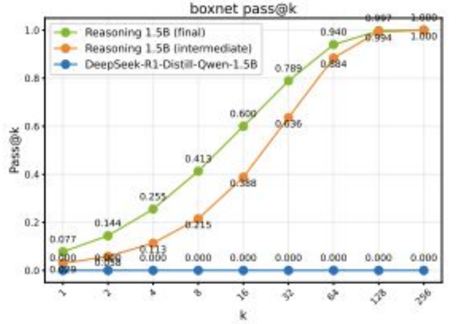
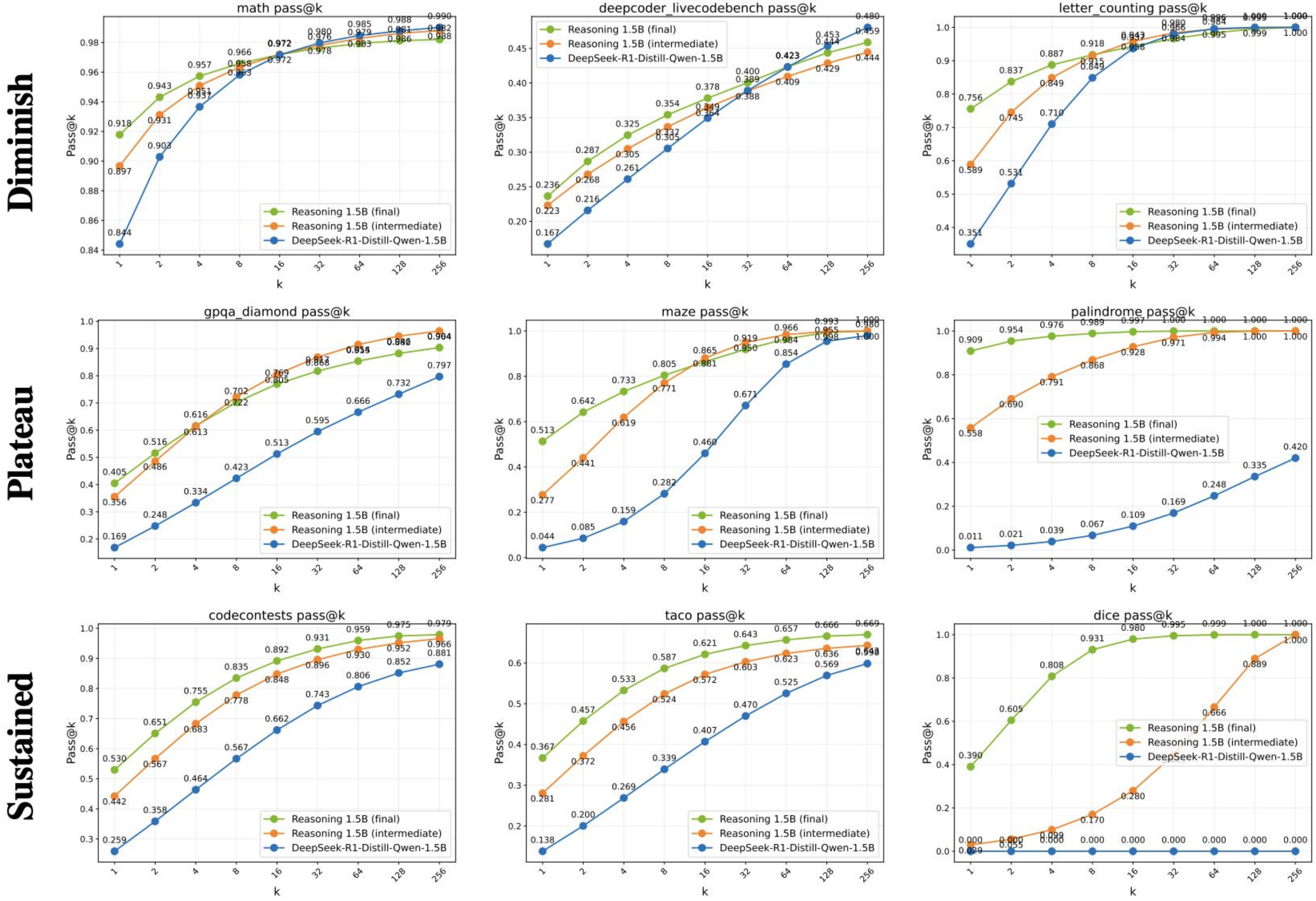
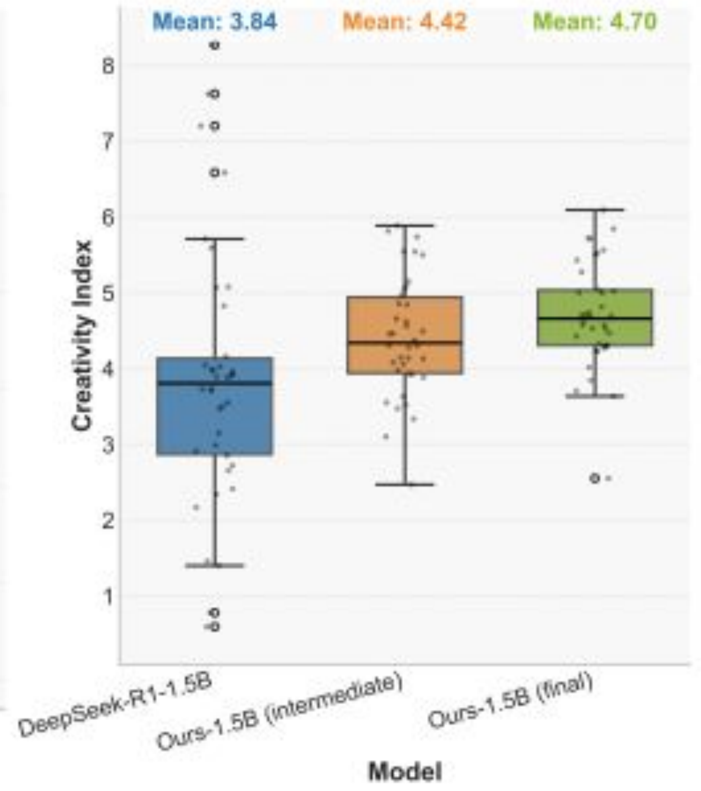
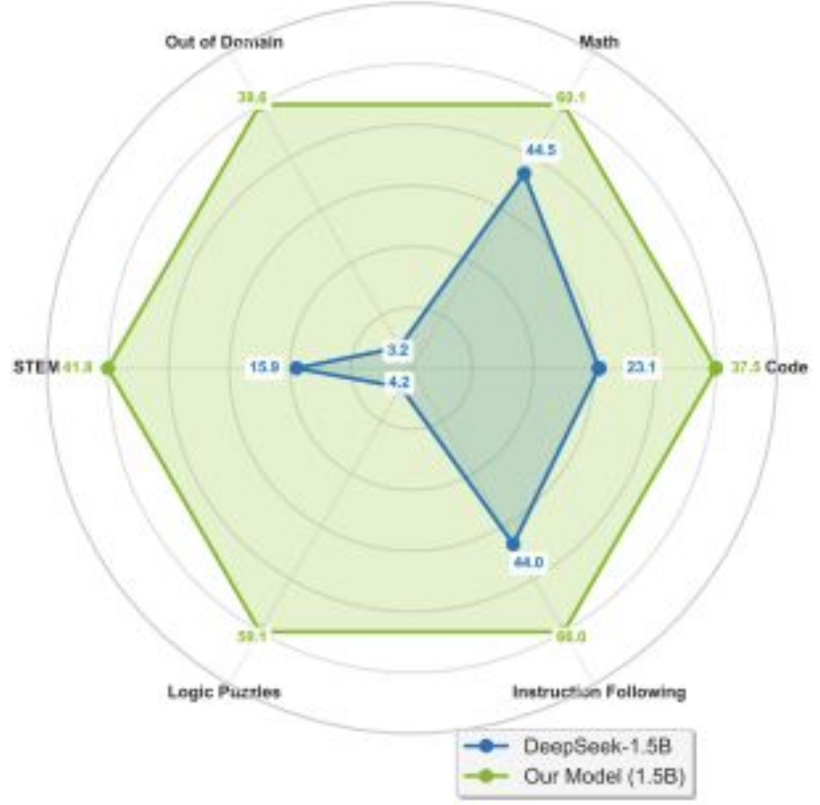


Figure 5: Expanded reasoning boundary for OOD task *boxnet*.


Diminish
Plateau
Sustained
math pass@k
0.918 0.943 0.957 0.966 0.972 0.980 0.985 0.988 0.990
0.897 0.931 0.951 0.958 0.972 0.976 0.979 0.981 0.982
0.844 0.903 0.937 0.963 0.972 0.978 0.983 0.986 0.988
deepcoder_livecodebench pass@k
0.236 0.287 0.325 0.354 0.378 0.400 0.423 0.453 0.480
0.223 0.268 0.305 0.337 0.349 0.389 0.409 0.429 0.444
0.167 0.216 0.261 0.305 0.364 0.388 0.423 0.444 0.459
letter_counting pass@k
0.756 0.837 0.887 0.918 0.943 0.980 0.996 1.000 1.000
0.589 0.745 0.849 0.915 0.937 0.966 0.984 0.999 1.000
0.351 0.531 0.710 0.849 0.958 0.984 0.995 0.999 1.000
gpqa_diamond pass@k
0.405 0.516 0.616 0.702 0.769 0.817 0.855 0.882 0.904
0.356 0.486 0.613 0.722 0.805 0.868 0.914 0.946 0.964
0.169 0.248 0.334 0.423 0.513 0.595 0.666 0.732 0.797
maze pass@k
0.513 0.642 0.733 0.805 0.865 0.919 0.966 0.993 1.000
0.277 0.441 0.619 0.771 0.881 0.950 0.984 0.998 0.980
0.044 0.085 0.159 0.282 0.460 0.671 0.854 0.955 1.000
palindrome pass@k
0.909 0.954 0.976 0.989 0.997 1.000 1.000 1.000 1.000
0.558 0.690 0.791 0.868 0.928 0.971 0.994 1.000 1.000
0.011 0.021 0.039 0.067 0.109 0.169 0.248 0.335 0.420
codecontests pass@k
0.530 0.651 0.755 0.835 0.892 0.931 0.959 0.975 0.979
0.442 0.567 0.683 0.778 0.848 0.896 0.930 0.952 0.966
0.259 0.358 0.464 0.567 0.662 0.743 0.806 0.852 0.881
taco pass@k
0.367 0.457 0.533 0.587 0.621 0.643 0.657 0.666 0.669
0.281 0.372 0.456 0.524 0.572 0.603 0.623 0.636 0.643
0.138 0.200 0.269 0.339 0.407 0.470 0.525 0.569 0.598
dice pass@k
0.390 0.605 0.808 0.931 0.980 0.995 0.999 1.000 1.000
0.029 0.055 0.099 0.170 0.280 0.666 0.889 1.000
0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
Pass@k
k
1 2 4 8 16 32 64 128 256
Reasoning 1.5B (final)
Reasoning 1.5B (intermediate)
DeepSeek-R1-Distill-Qwen-1.5B


Mean: 3.84
Mean: 4.42
Mean: 4.70
Creativity Index
1
2
3
4
5
6
7
8
DeepSeek-R1-1.5B
Ours-1.5B (intermediate)
Ours-1.5B (final)
Model


Out of Domain
Math
Code
Instruction Following
Logic Puzzles
STEM
39.6
60.1
37.5
66.0
59.1
41.8
44.5
23.1
44.0
4.2
15.9
3.2
DeepSeek-1.5B
Our Model (1.5B)

Conclusion

- **Conclusive Results** : Stable and prolonged RL training expands reasoning capabilities.
- **Effective RL Training Requires** : Long training, reference resets, and diverse tasks.
- **Future Impact** : Small models can achieve powerful reasoning, democratizing high-level AI access.

Limitations

- **Limited Task Scope:** Small subset of all tasks
- **Scalability Concerns:** Work done on smaller models
- **Computational Resources:** Extensive Training = Great Computational Demands
- **Unstable Training:** Hard Resets leads to variability in repeated trials

Culminating Conclusion

- **RLVR works:** big gains on verifiable reasoning tasks (math/code)
- But “reasoning gains” could either mean efficiency or capability
- **DeepSeek-R1:** RLVR can elicit longer, self-correcting reasoning behavior
- **Boundary paper:** boosts pass@1 but may not expand and can even shrink reasoning boundaries
- **ProRL:** prolonged/stabilized RL can expand reasoning boundaries

Main Takeaway: RLVR is rapidly evolving; results depend on the training pipeline, and new research is constantly improving outcomes

Thank you. Any Questions?