DeepMath-103K: A Large-Scale, Challenging, Decontaminated, and Verifiable Mathematical Dataset for Advancing Reasoning

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang,
Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang,
Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, Dong Yu

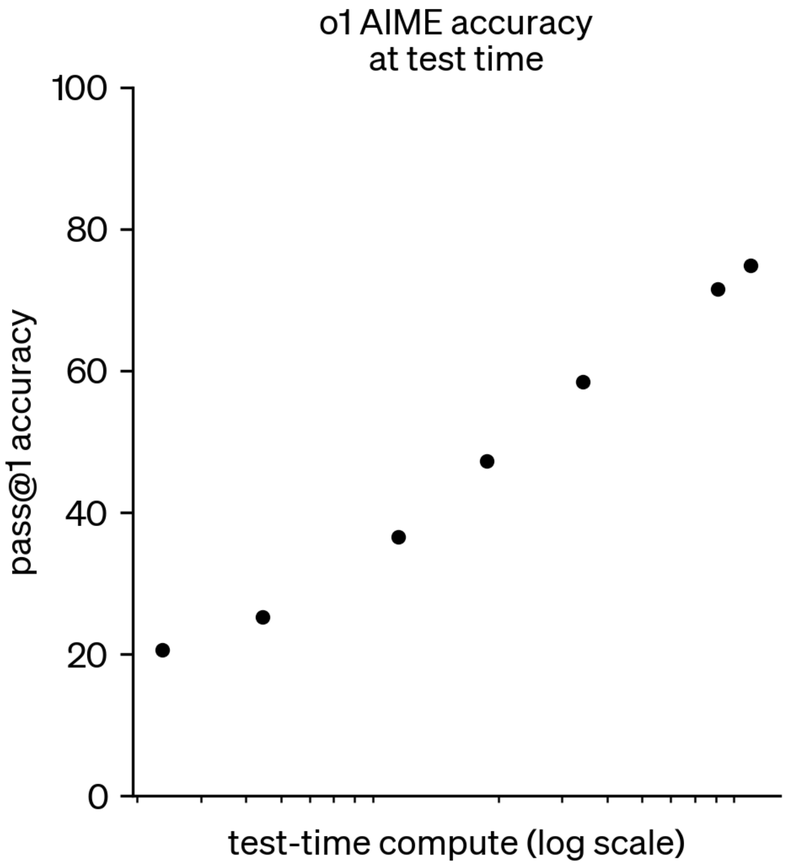
🤗 Data: https://hf.co/datasets/zwhe99/DeepMath-103K
🤗 Model: https://hf.co/collections/zwhe99/deepmath-6816e139b7f467f21a459a9a
💻 Code: https://github.com/zwhe99/DeepMath

Presenter: Zhiwei He

Background: scaling test-time compute

OpenAI-o1 (2024.09)

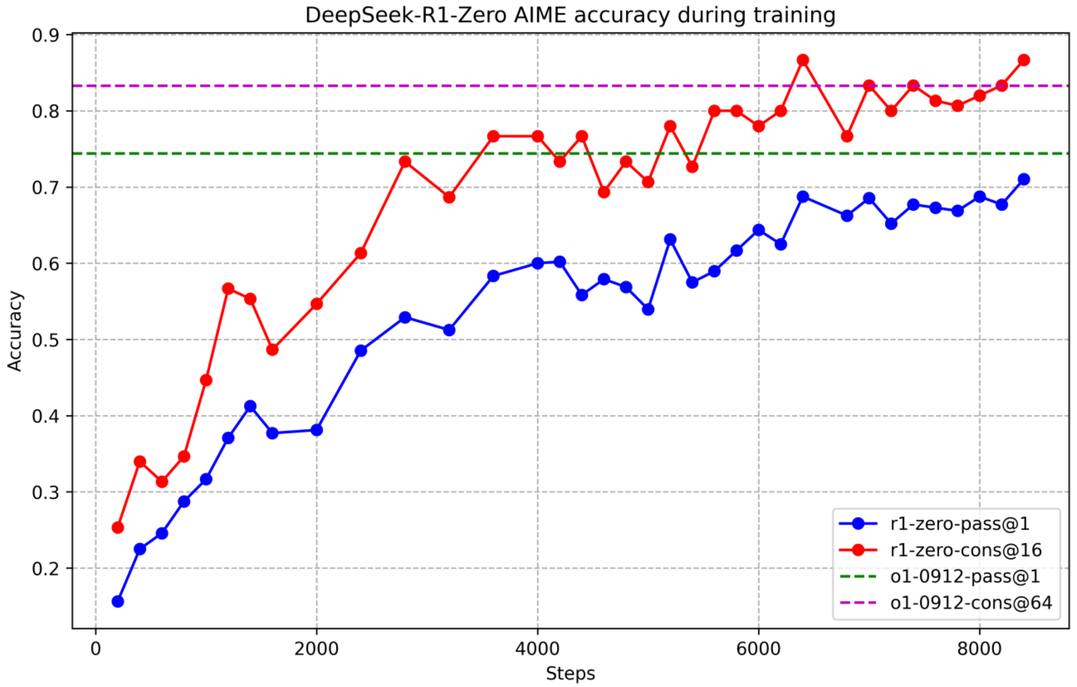
- The performance of o1 consistently improves with more time spent thinking (test-time compute).
- Human-like thinking patterns
  - exploring multiple strategies
  - breaking down complex steps
  - double-checking
  - self-reflection
  - …

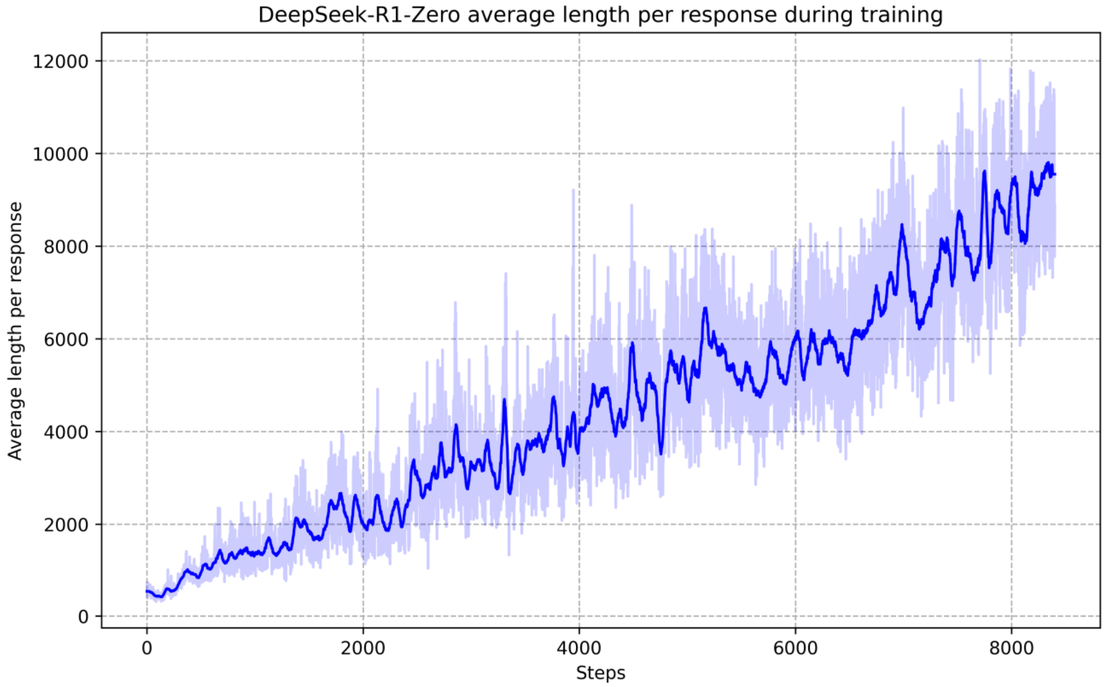


[1] Learning to reason with LLMs (https://openai.com/index/learning-to-reason-with-llms)

Background: scaling test-time compute

DeepSeek-R1 (2025.01)

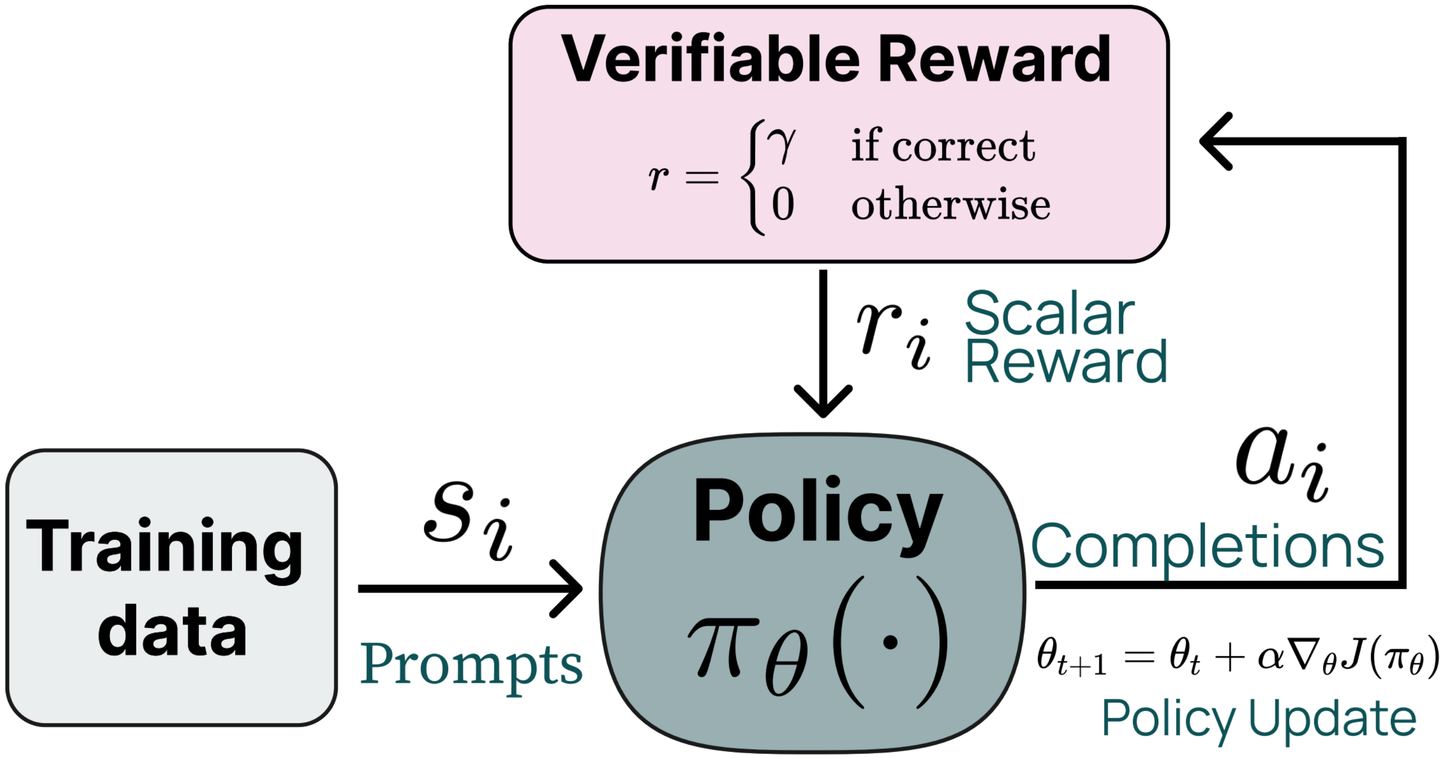




[2] DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

Background: Reinforcement Learning with Verifiable Rewards

RLVR



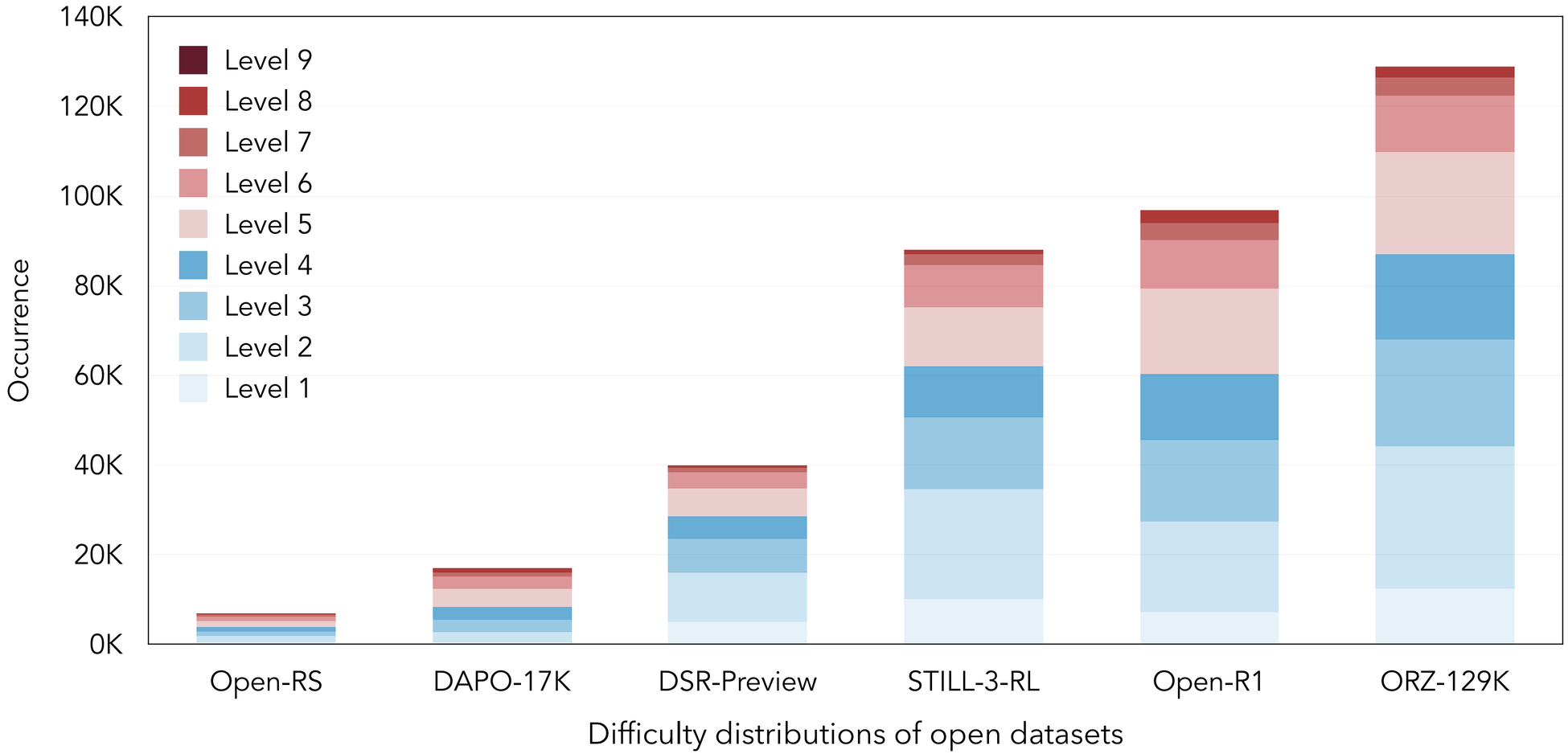
[3] Tulu 3: Pushing Frontiers in Open Language Model Post-Training

Efforts of the open-source community

- Recipe
  - SimpleRL
  - Open-Reasoner-Zero
  - …
- Algorithm
  - DAPO
  - Dr GRPO
  - …
- Data
  - ORZ-129K
  - Open-R1
  - STILL-3-RL-90K
  - DeepScaleR-Preview
  - DAPO-17K

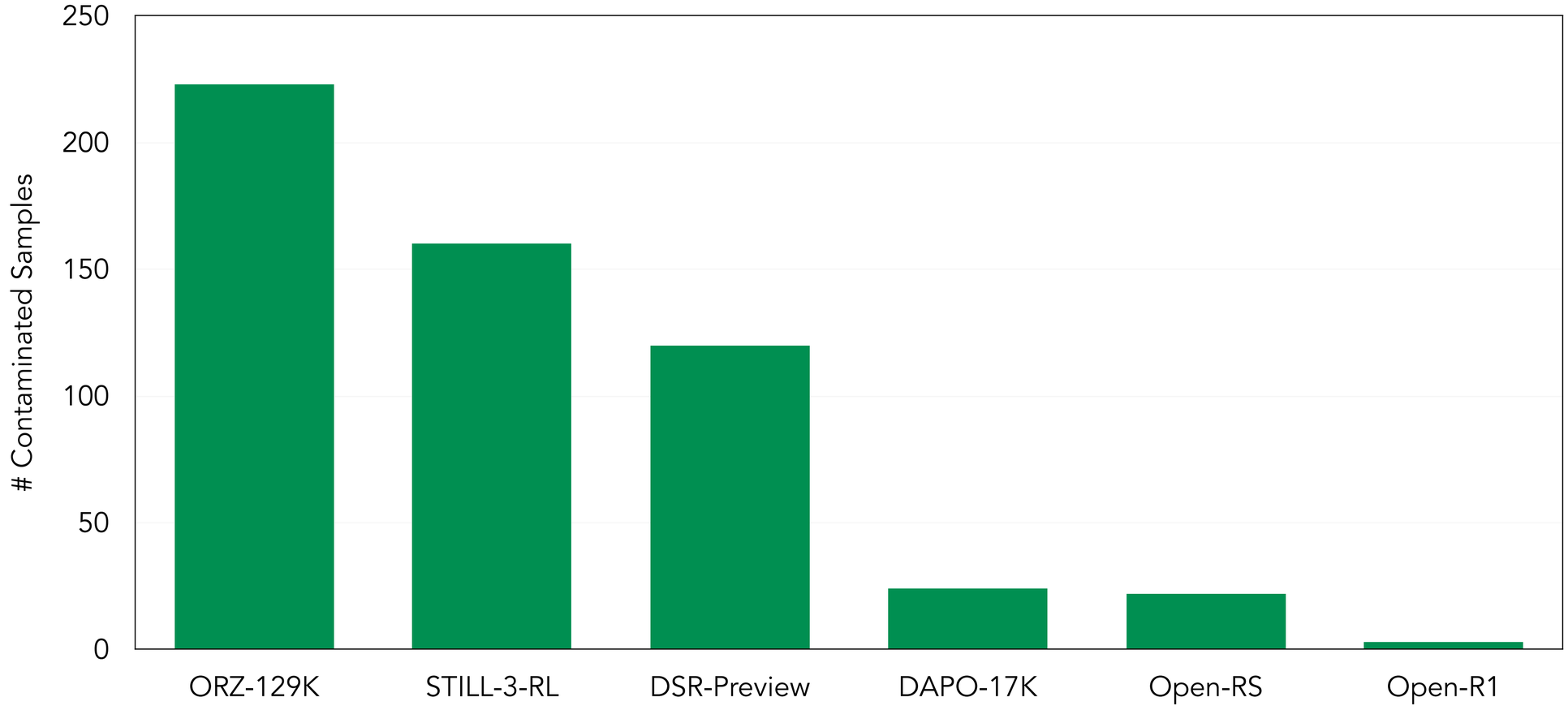
Existing open datasets for RLVR

Insufficient difficulty



Existing open datasets for RLVR

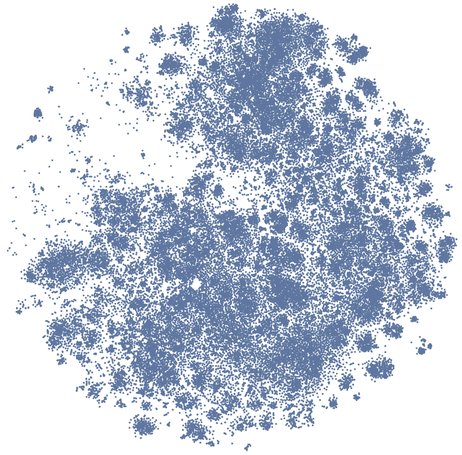
Being contaminated with benchmarks



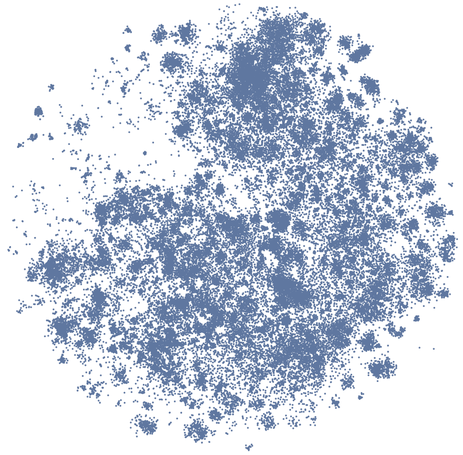
Many open datasets exhibit some degree of contamination with MATH500 (string-based method)

Existing open datasets for RLVR

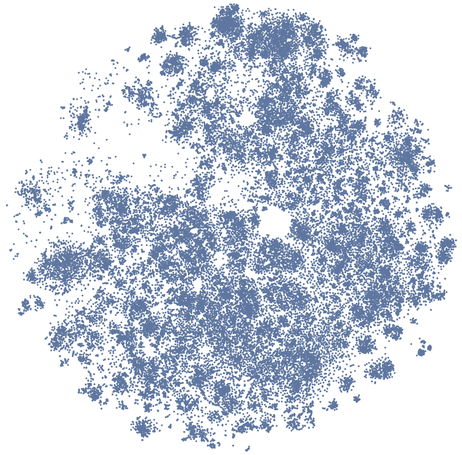
Overlapping with each other



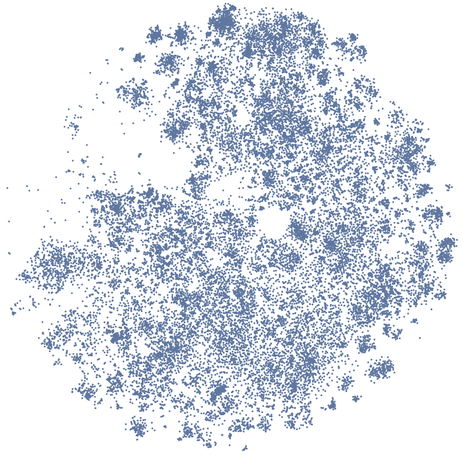
ORZ-129K



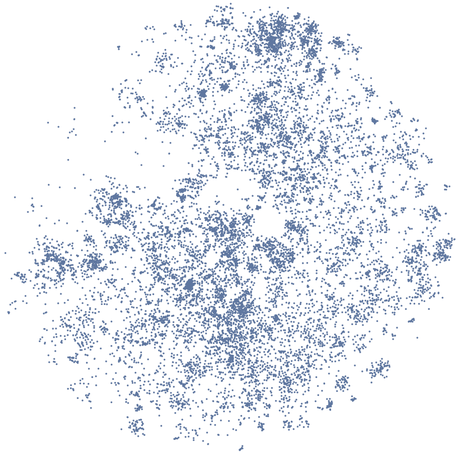
Open-R1



STILL-3-RL-90K



DSR-Preview



DAPO-17K

DAPO-17K

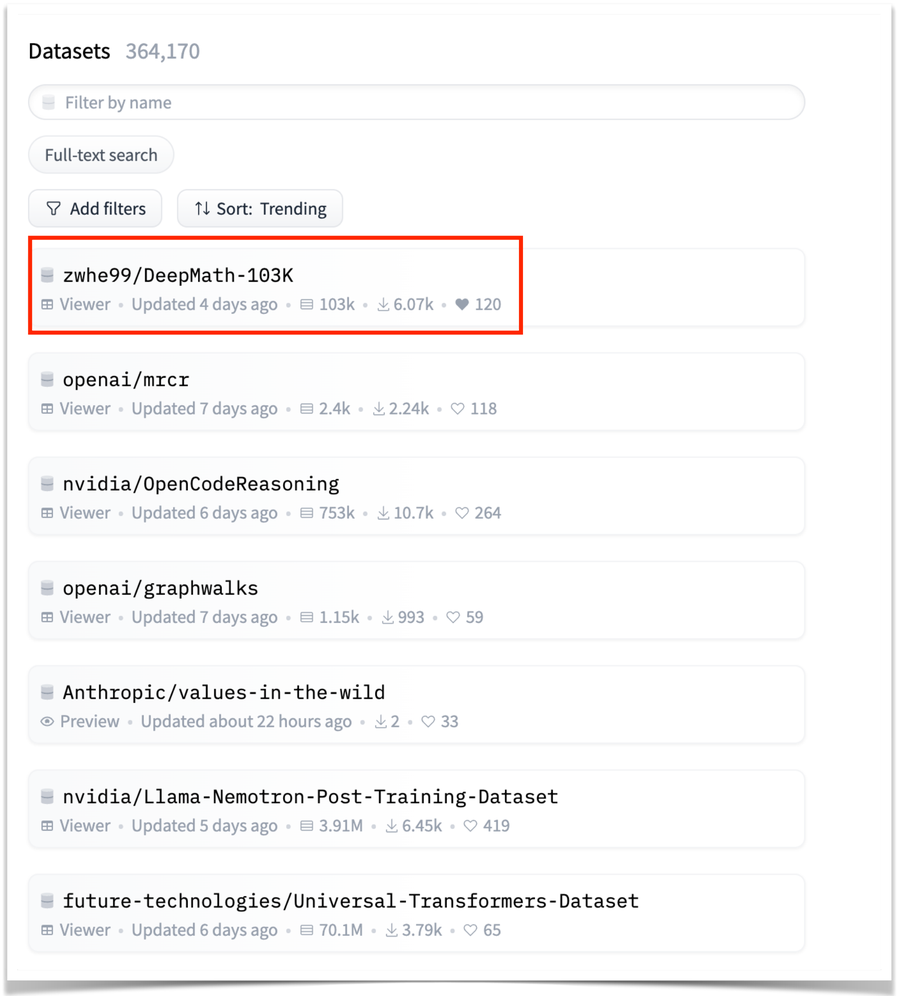
DeepMath-103K

Our goals

- Challenging
- Decontaminated
- Verifiable
- Novel

DeepMath-103K

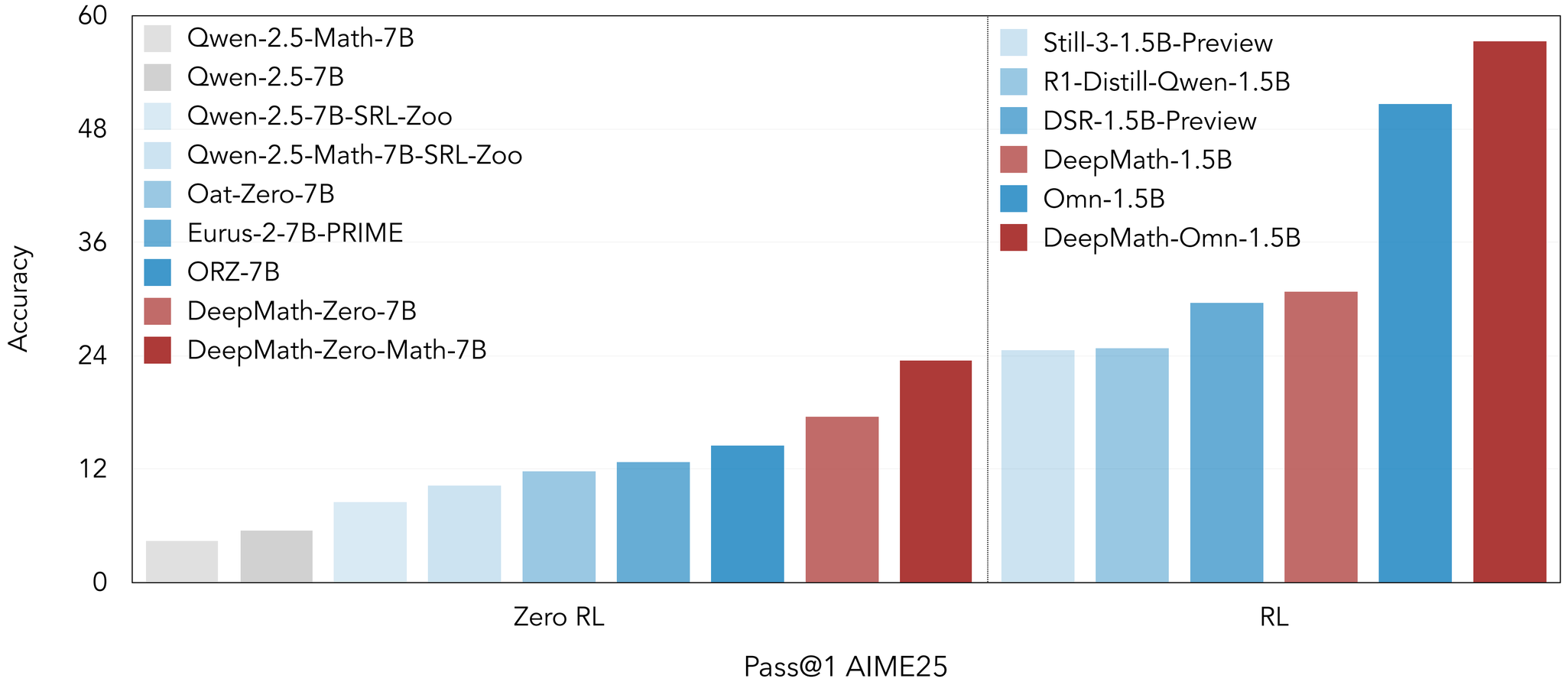
Top trending dataset @ HuggingFace



2025.05.28

DeepMath Series Modes

State-of-the-art performance



Overview of DeepMath-103K

A data sample from DeepMath-103K

Question: Calculate the line integral $\oint_C P\,dx + Q\,dy$, over the ellipse $\frac{x^2}{25} + \frac{y^2}{36} = 1$, where the vector fields are given by: $P = \frac{-y}{(x-1)^2 + y^2}$, $Q = \frac{x-1}{(x-1)^2 + y^2}$. Determine the value of the integral, considering that the vector field is undefined at the point $(0,1)$ inside the ellipse.

Final Answer: 2π

Difficulty: 8

Topic: Mathematics -> Calculus -> Integral Calculus -> Techniques of Integration -> Multi-variable

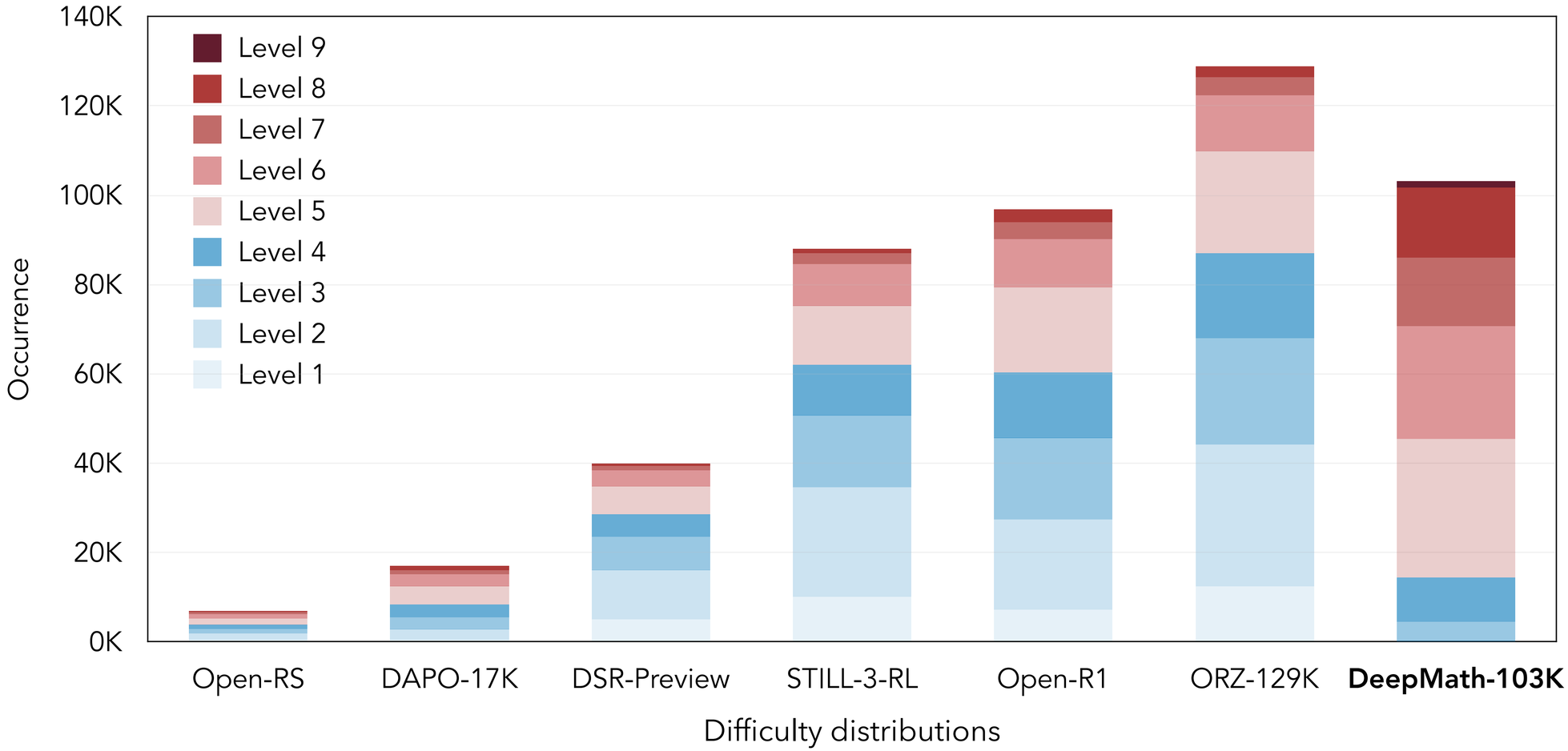
R1 Solution 1: Okay, so I need to calculate the line integral … Hmm, the problem also mentions that … Thus, the value of the line integral is: 2π

R1 Solution 2: Okay, so I need to calculate the line integral …. Hmm, first things first, let me recall what line integrals are about … Thus, the value of the line integral is: 2π

R1 Solution3: Okay, so I need to calculate the line integral … So, first, maybe I should visualize the ellipse … Thus, the value of the line integral is: 2π

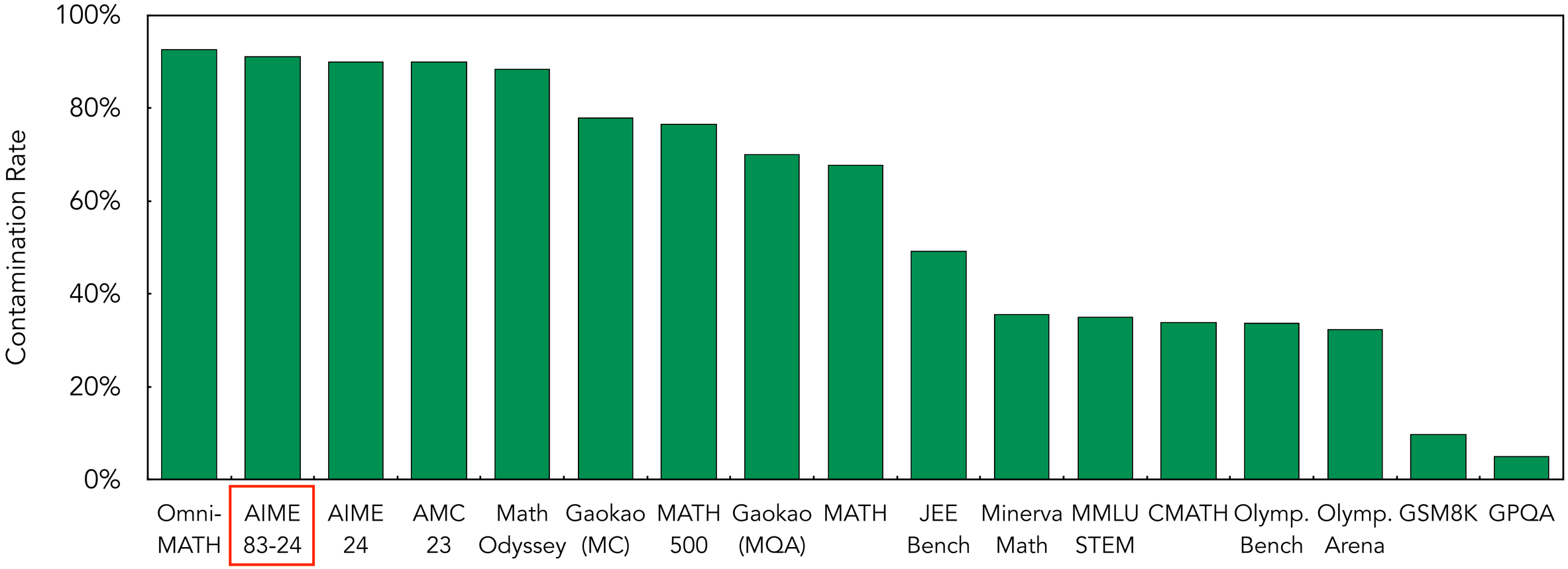
Overview of DeepMath-103K

DeepMath-103K is challenging compared to existing datasets



Overview of DeepMath-103K

Rigorous Data Decontamination

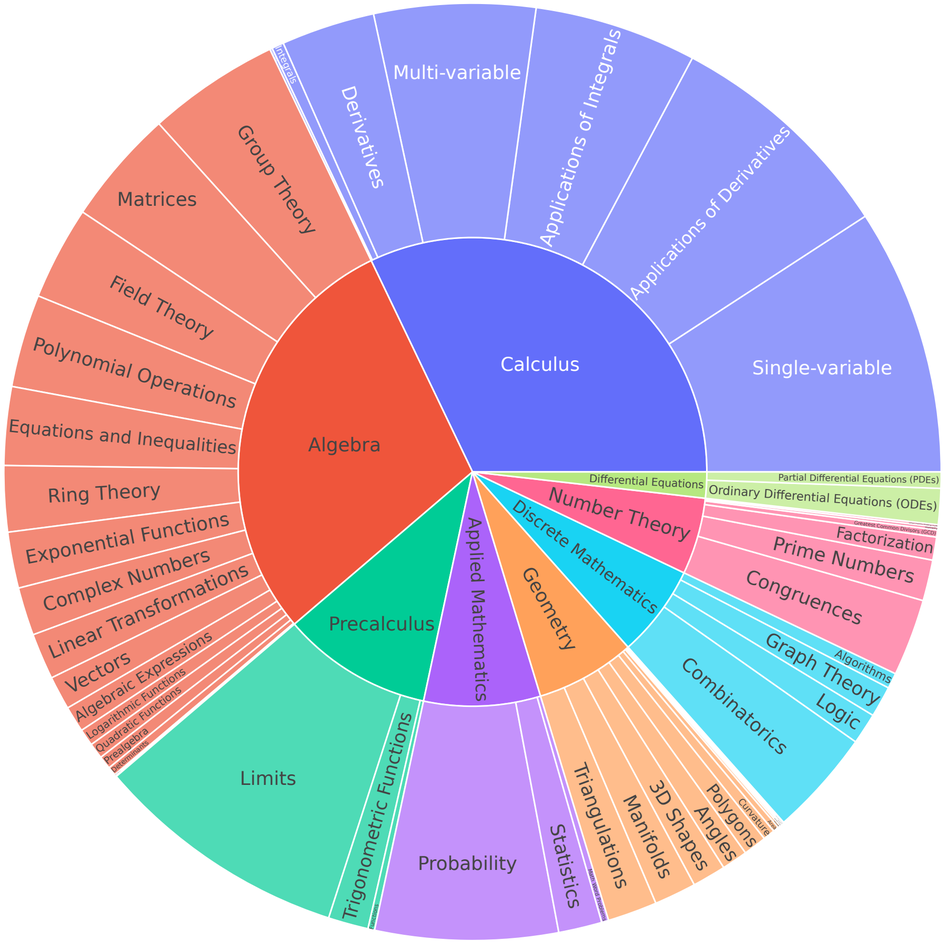


Contamination rates of common mathematical and STEM benchmarks detected in the raw data sources before decontamination.

Overview of DeepMath-103K

Broad Topical Diversity

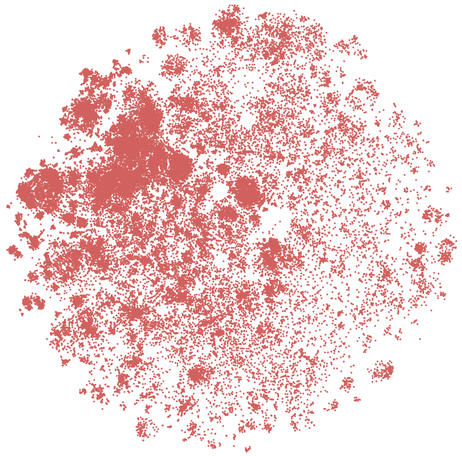
- Fundamental topics
  - Prealgebra
  - Plane Geometry
  - …
- Sophisticated topics
  - Abstract Algebra
  - Advanced Calculus
  - …



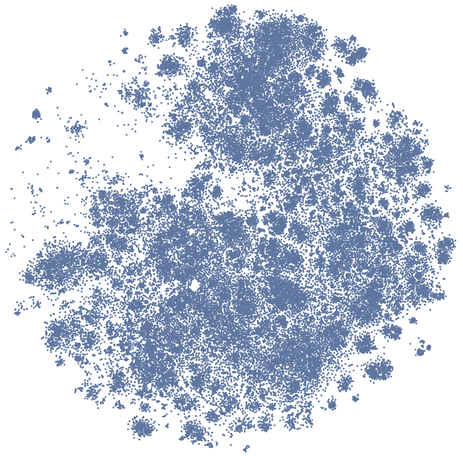
Hierarchical breakdown of covered mathematical topics in DeepMath-103K.

Overview of DeepMath-103K

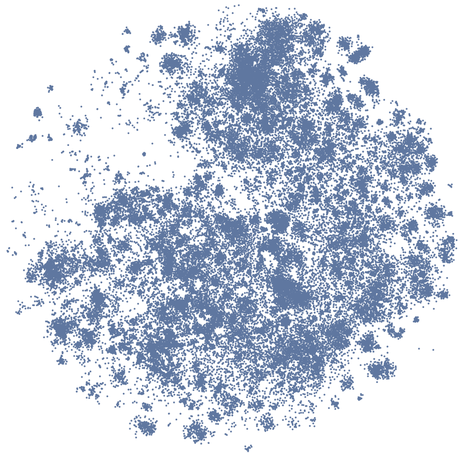
Data Novelty and Uniqueness



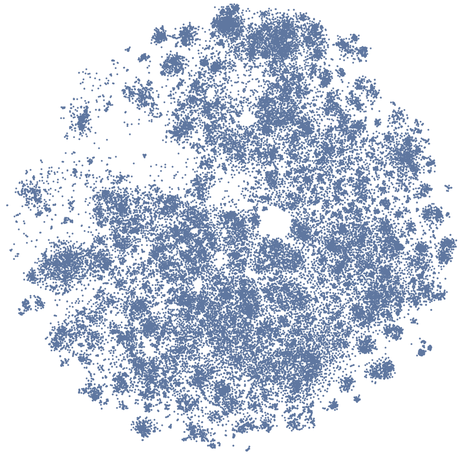
DeepMath-103K



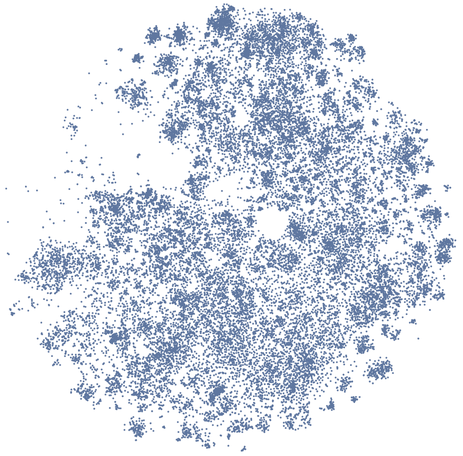
ORZ-129K



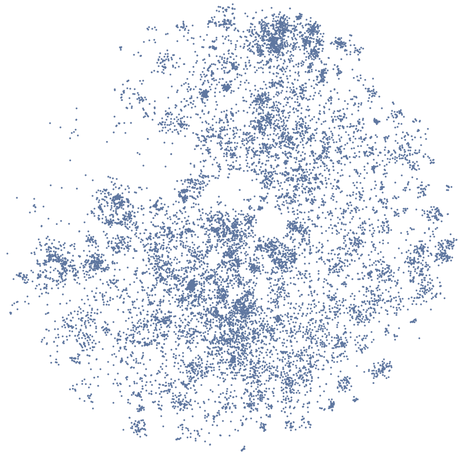
Open-R1



STILL-3-RL-90K



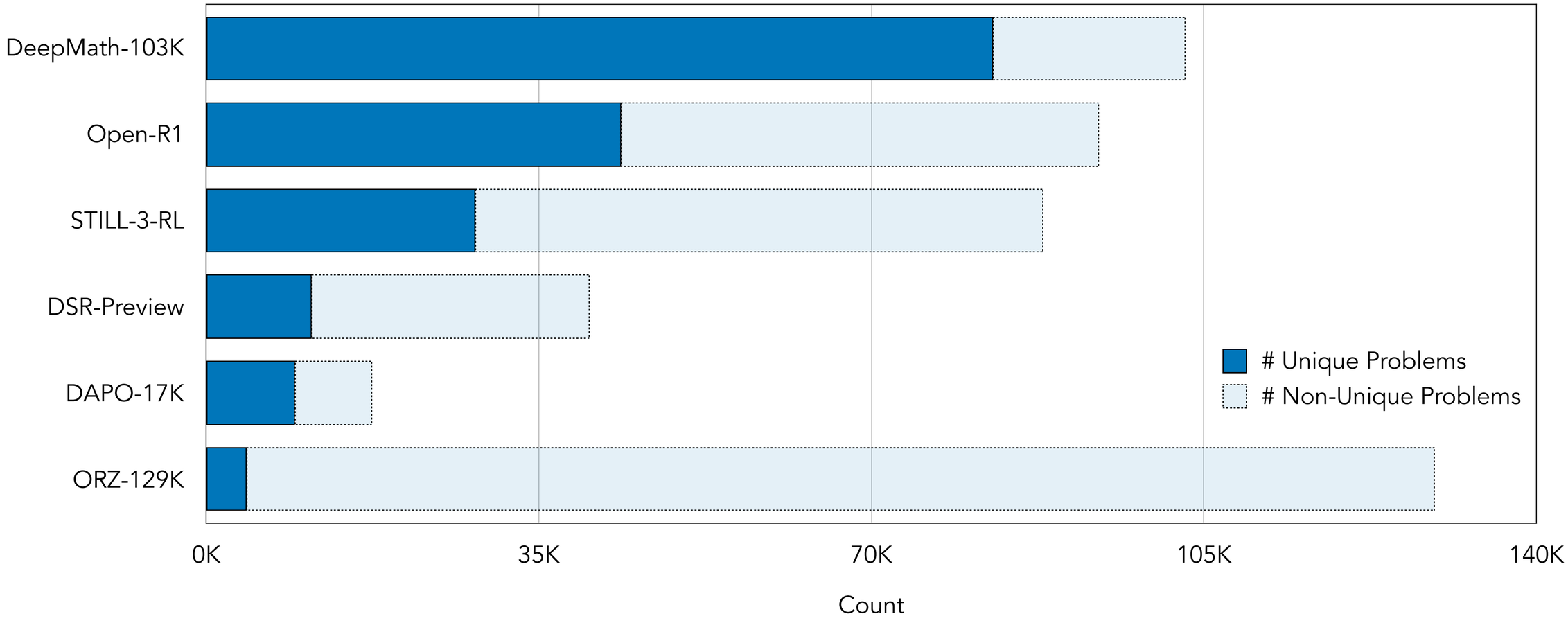
DSR-Preview



DAPO-17K

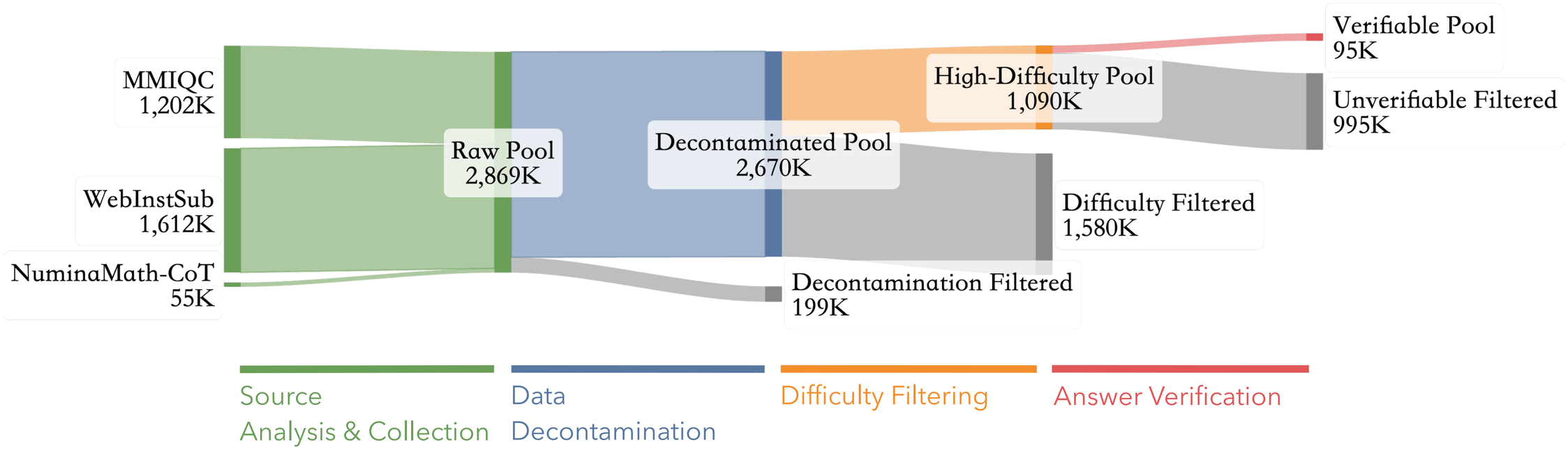
Overview of DeepMath-103K

Data Novelty and Uniqueness



Unique and non-unique problem counts in DeepMath-103K compared to other datasets.

Construction of DeepMath-103K

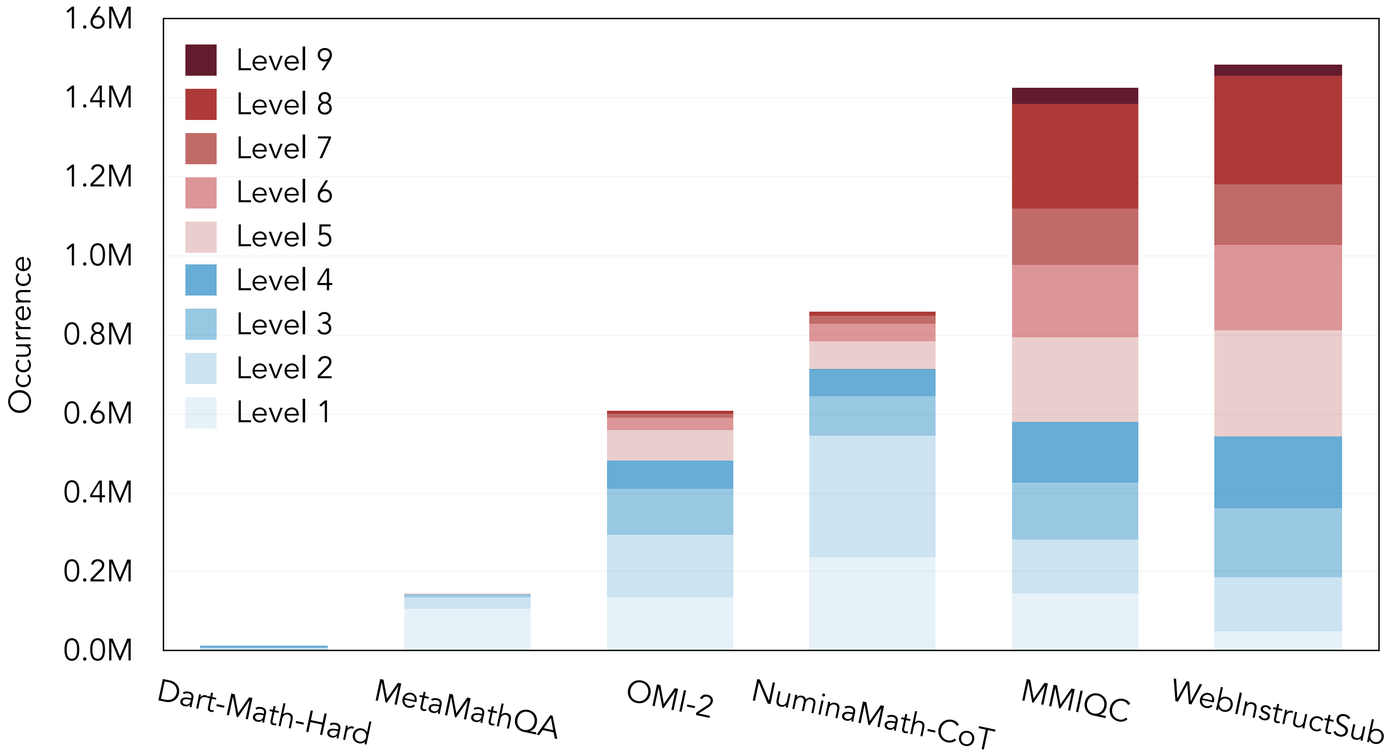


138,000 US dollars & 127,000 H20 GPU hours

Construction of DeepMath-103K

Source Analysis & Collection

- We considered 6 SFT datasets as potential sources.
  - Dart-Math-Hard, MetaMathQA and OMI-2 augment problems and solutions from GSM8K and MATH.
  - NuminaMath-CoT, MMIQC, and WebInstructSub source content more broadly from the web.
- High-difficulty problems are almost all from Math Stack Exchange.
- Final sources
  - NuminaMath-CoT (olymp)
  - MMIQC (Math Stack Exchange)
  - WebInstructSub (Math Stack Exchange)



Difficulty distributions of various open mathematical datasets considered as potential sources.

Construction of DeepMath-103K

Data Decontamination

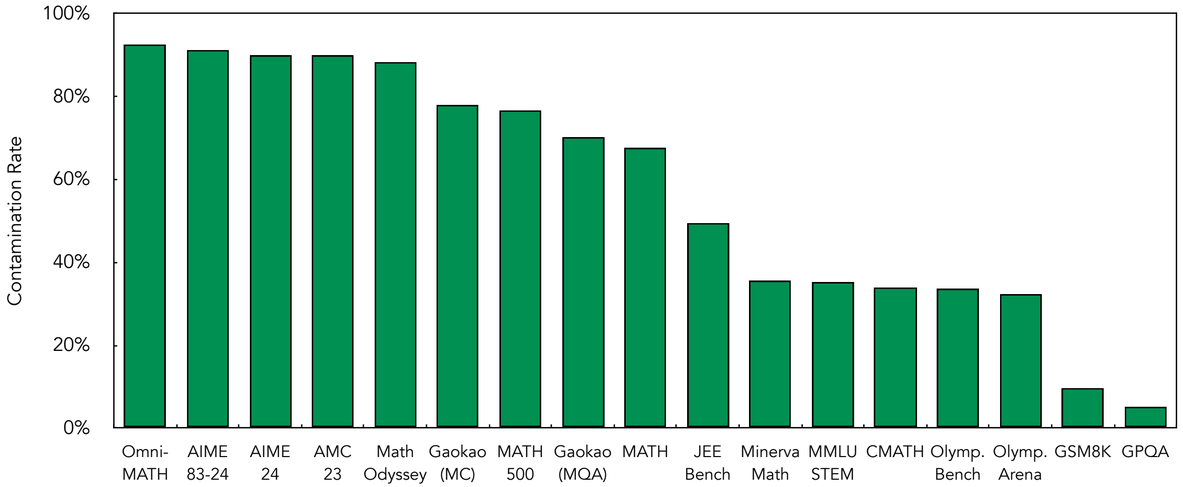
- For each candidate question in our raw dataset:

1. **Semantic Search**: find the top-5 most similar examples from all aggregated test sets using embedding similarity search.

2. **LLM Judge**: ask an LLM-Judge if any of these 5 pairs should be considered as identical questions or paraphrases.

Benchmark	Raw Question	Benchmark Question
AIME24	How many routes are there through from top left corner to bottom right in a 20x20 grid? I'm trying to solve this computer programming problem on Project Euler. I've seen a solution using nCr, where n = 40 and r = 20. Could someone explain to me how this work, please?	Consider the paths of length 16 that follow the lines from the lower left corner to the upper right corner on an 8x8 grid. Find the number of such paths that change direction exactly four times, as in the examples shown below.
AMC23	Using only 3 paise, 5 paise, and 9 paise coins, what is the largest amount that cannot be paid in exact change?	In the state of Coinland, coins have values 6,10, and 15 cents. Suppose x is the value in cents of the most expensive item in Coinland that cannot be purchased using these coins with exact change. What is the sum of the digits of x?

Examples of detected contamination. Colors highlight similarities.



Contamination rates of common benchmarks detected in the raw data sources before decontamination.

Construction of DeepMath-103K

Difficulty Filtering

1. We assigned a difficulty level to each decontaminated problem by prompting GPT-4o based on the annotation guidelines provided by the AoPS [4].

2. To ensure a robust estimate, we queried GPT-4o **six** times for each problem and averaged the resulting ratings to determine its final difficulty level.

3. Level 5-9 are retained.

Table 2: Examples of geometry problems retained by the difficulty filtering process (level $\geqslant 5$).

Difficulty	Problem
5	Four random points are placed in the plane, with each point's horizontal and vertical coordinates uniformly distributed on the interval $(0, 1)$. What is the expected largest size of a subset of these points that can form the vertices of a convex polygon?
6	A square has one side lying on the line $y = 2x - 17$ and two other vertices on the parabola $y = x^2$. Determine the minimum possible area of the square."
7	Determine the sequence $s(k, n)$, which represents the number of sides of the intersection of a unit-radius regular polygon P_k with k sides and a rotating unit-radius regular polygon P_n with $n \geqslant k$ sides, as the angle of rotation θ varies from 0 to 2π. Provide the sequence $s(k, n)$ for all $n \geqslant k$.
8	Consider a convex n-gon $A_1A_2 \cdots A_n$ inscribed in a unit circle. Determine the maximum value of the sum of the squares of all its sides and diagonals
9	Determine the maximal cardinality of a collection $\mathcal{C}$ of projective planes on ω such that no two distinct members of $\mathcal{C}$ are isomorphic. A set $L \subseteq \mathcal{P}(X)$ is a projective plane on $X \neq \varnothing$ if: 1. For any distinct $x, y \in X$, there is a unique $l \in L$ such that $x, y \in l$. 2. For any distinct $l, m \in L$, $\lvert l \cap m \rvert = 1$. 3. There exist four distinct elements of X such that no member of L contains more than two of these four elements. Two projective planes L and M on X are isomorphic if there is a bijection $\varphi : X \to X$ such that $l \in L$ if and only if $\varphi(l) \in M$.

[4] https://artofproblemsolving.com/wiki/index.php/AoPS%20Wiki:Competition%20ratings

Construction of DeepMath-103K

Answer Verification

- Two main challenges:
  1. Open-ended questions inherently lack a easily verifiable final answer.
  2. Some answers are excessively complex (e.g., lengthy expressions or intricate notation), making them challenging or even infeasible for automated rule-based verification.

Can you explain the intuition behind the promotion of the second BC inference, which replaces the pairwise independence condition with the weaker condition $P(A_k A_j) \leq P(A_k)P(A_j)$ (for all k and j, where $k \neq j$)? How is this condition (called pairwise negative correlation) different from independence? How does it affect the results?"

An open-ended question

$\bigoplus_{i=0}^{p_1-1} \mathbb{Z}_{p_2}$ 和 $\bigoplus_{k=0}^{p_1-1} \mathbb{Z}_{p_2}$

$\bigoplus_{i=0}^{p_1-1} \mathbb{Z}_{p_2}[\theta_{p_1}^i]$, $\mathbb{Z}_{p_2} \oplus \bigoplus_{k=1}^{p_1-1} \mathbb{Z}_{p_2}[\theta_{p_1}^k]$, $\bigoplus_{d|p_1} \mathbb{Z}_{p_2}[\theta_d]$

$$\frac{1}{z-2} + \frac{1-2i}{5} \sum_{n=0}^{\infty} \left(\frac{z-2}{2-i}\right)^n + \frac{1+2i}{5} \sum_{n=0}^{\infty} \left(\frac{z-2}{2+i}\right)^n$$

Lengthy expressions and intricate notation

[4] https://artofproblemsolving.com/wiki/index.php/AoPS%20Wiki:Competition%20ratings

Construction of DeepMath-103K

Answer Verification

- Two-stage verification process:
  1. **Question Filtering and Formatting using GPT-4o**:
     - Problem types inherently unsuitable for verification were discarded.
     - Questions phrased conversationally were rewritten into a standardized format seeking a single, specific numerical or symbolic answer.
  2. **Answer Verification via Consistency Check**:
     - We generated three distinct solution paths using DeepSeek-R1 for each question.
     - A rule-based verifier then extracted the final answer from each of these generated solutions, as well as from the original source solution (when available).
     - Only problems where all extracted final answers were identical were retained in the final dataset.

[4] https://artofproblemsolving.com/wiki/index.php/AoPS%20Wiki:Competition%20ratings

DeepMath Series Models

Math reasoning performance

Table 3: Math reasoning performance. "DeepMath" denotes models trained on DeepMath-103K.

Model	MATH 500	AMC 23	Olympiad Bench	Minerva Math	AIME 24	AIME 25	Poly Math
Proprietary Models							
o1-mini	–	–	–	–	63.6	–	–
o3-mini (low effort)	–	–	–	–	60.0	–	–
Zero RL from Base Model							
Qwen-2.5-7B (Team, 2024)	54.8	35.3	27.8	16.2	7.7	5.4	28.1
↳ Open-Reasoner-Zero-7B (Hu et al., 2025)	81.8	58.9	47.9	38.4	15.6	14.4	40.7
↳ Qwen-2.5-7B-SRL-Zoo (Zeng et al., 2025a)	77.0	55.8	41.0	41.2	15.6	8.7	33.1
↳ DeepMath-Zero-7B (Ours)	**85.5**	**64.7**	**51.0**	**45.3**	**20.4**	**17.5**	**42.7**
Qwen-2.5-Math-7B (Team, 2024)	46.9	31.9	15.8	15.5	11.2	4.4	22.7
↳ Qwen-2.5-Math-7B-SRL-Zoo (Hu et al., 2025)	75.8	59.7	37.4	29.9	24.0	10.2	36.0
↳ Qat-Zero-7B (Liu et al., 2025)	80.0	66.7	43.4	40.8	32.7	11.7	40.8
↳ Eurus-2-7B-PRIME (Cui et al., 2025)	80.2	64.7	44.9	42.1	19.0	12.7	38.9
↳ DeepMath-Zero-Math-7B (Ours)	**86.9**	**74.7**	**52.3**	**49.5**	**34.2**	**23.5**	**46.6**
RL from Instruct Models							
R1-Distill-Qwen-1.5B (Guo et al., 2025)	84.7	72.0	53.1	36.6	29.4	24.8	39.9
↳ DeepScaleR-1.5B-Preview (Luo et al., 2025)	89.4	80.3	60.9	42.2	**42.3**	29.6	**46.8**
↳ Still-3-1.5B-Preview (Chen et al., 2025)	86.6	75.8	55.7	38.7	30.8	24.6	43.1
↳ DeepMath-1.5B (Ours)	**89.9**	**82.3**	**61.8**	**42.5**	37.3	**30.8**	46.6
OpenMath-Nemotron-1.5B (Moshkov et al., 2025)	91.8	90.5	70.3	26.3	61.3	50.6	56.8
↳ DeepMath-Omn-1.5B (Ours)	**93.2**	**94.2**	**73.4**	**28.3**	**64.0**	**57.3**	**58.7**

DeepMath Series Models

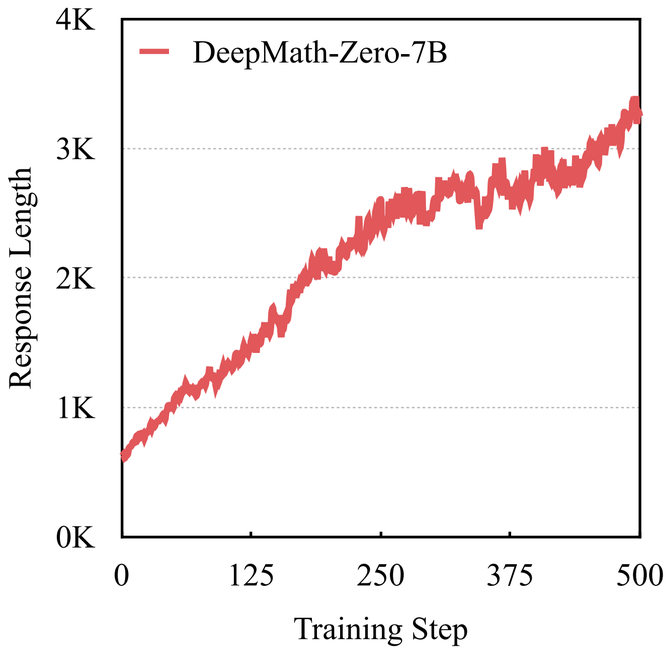
Generalizable Reasoning beyond Mathematics

Table 4: Performance on the GPQA-Diamond benchmark.

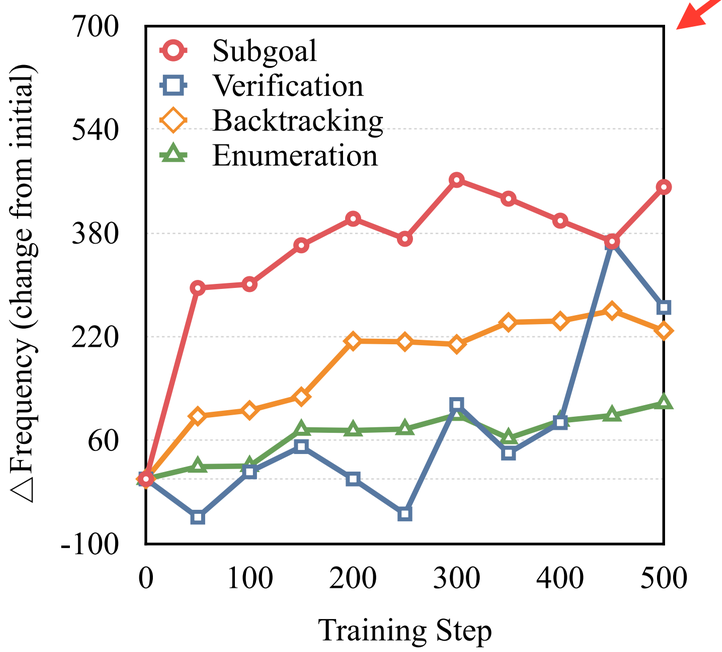
Model	Biology	Physics	Chemistry	Overall
Zero RL from Base Model				
Qwen-2.5-7B	33.6	27.8	21.4	25.3
↳ Open-Reasoner-Zero-7B	50.3	47.8	26.7	38.1
↳ Qwen-2.5-7B-SimpleRL-Zoo	31.9	37.9	22.6	30.2
↳ DeepMath-Zero-7B (Ours)	**57.2**	**53.0**	**28.2**	**41.7**
Qwen-2.5-Math-7B	32.2	26.0	21.1	24.3
↳ Qwen-2.5-Math-7B-SRL-Zoo	40.1	31.2	22.9	28.2
↳ Qat-Zero-7B	**49.0**	36.8	22.0	31.0
↳ Eurus-2-7B-PRIME	44.1	37.4	24.1	31.8
↳ DeepMath-Zero-Math-7B (Ours)	47.4	**56.3**	**26.0**	**41.2**
RL from Instruct Models				
R1-Distill-Qwen-1.5B	13.5	36.2	4.4	19.1
↳ DeepScaleR-1.5B-Preview	15.5	46.8	9.1	26.1
↳ Still-3-1.5B-Preview	16.8	38.4	5.2	20.7
↳ DeepMath-1.5B (Ours)	**18.1**	**47.6**	**12.2**	**28.2**
OpenMath-Nemotron-1.5B	12.8	23.5	18.9	20.3
↳ DeepMath-Omn-1.5B (Ours)	**17.1**	**28.4**	**21.5**	**24.1**

DeepMath Series Models

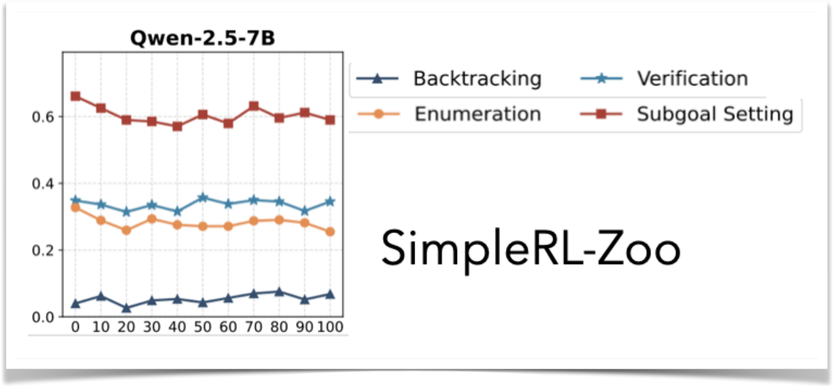
Analysis of Zero RL

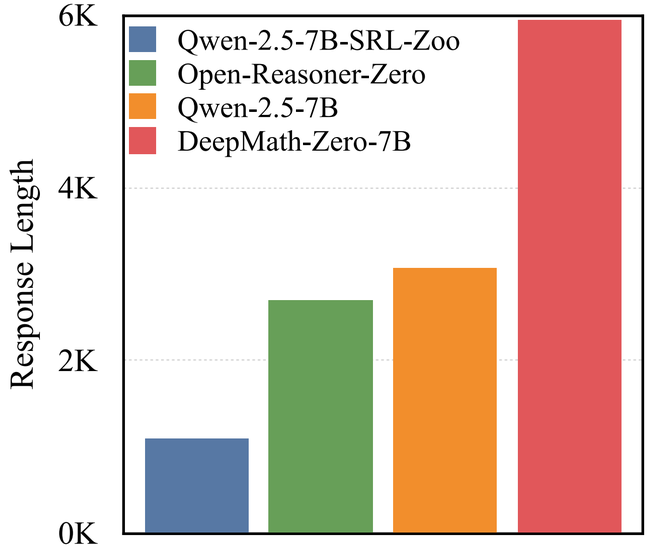


Rollout response length during training



Change of cognitive behaviors during training





Average response length on evaluated benchmarks

[5] SimpleRL-Zoo: Investigating and Taming Zero Reinforcement Learning for Open Base Models in the Wild

Summary

- We introduce DeepMath-103K, a large-scale mathematical dataset specifically designed to advance the reasoning capabilities of LLMs through RLVR.
- DeepMath-103K distinguishes itself through its high concentration of challenging problems, rigorous decontamination against a wide range of benchmarks, and the inclusion of verifiable final answers and multiple diverse solutions for each problem.
- The DeepMath series models achieve new SOTA results on many mathematical benchmarks and exhibit remarkable generalization to domains beyond mathematics.

Thank You