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

arxiv.org/abs/2504.11456

Zhiwei He*,1,2 Tian Liang*,1 Jiahao Xu*,1 Qiuzhi Liu[1] Xingyu Chen[1,2] Yue Wang[1]
Linfeng Song[1] Dian Yu[1] Zhenwen Liang[1] Wenxuan Wang[1] Zhuosheng Zhang[2]
Rui Wang†,2 Zhaopeng Tu†,1 Haitao Mi[1] Dong Yu[1]

[1]Tencent [2]Shanghai Jiao Tong University

https://github.com/zwhe99/DeepMath
https://hf.co/datasets/zwhe99/DeepMath-103K

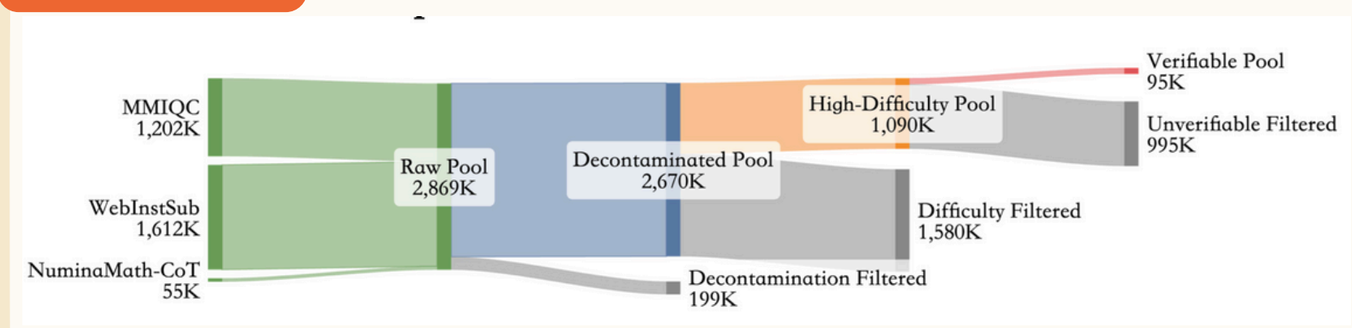
简介

本文提出DeepMath-103K，它是一个专为强化LLM数学推理能力而构建的大规模、高质量、有难度、题目新的数学数据集。与现有数学数据集相比，DeepMath的特点：难度高，如果将数学题难度划分为1-9级，DeepMath中的题目主要位于难度5-9；重合率低，作者并非对已有数据集整合，而是从Math StackExchange中筛选，有82K道新题目；元数据丰富，每道题都包含可验证的答案(可用于RLVR)、难度、主题、3个R1生成的solution(可用于sft)。

背景

数学推理可以说是LLM REASONING中最受关注的一个任务，现在学术界主流方法是RLVR，但是RLVR也不是万能的，它也需要高质量的训练集。而现有数学数据集普遍存在的第一个问题是难度偏低了，现在LLM能力越来越强，在很多数据集上效果准的飞起，如果追求排名的考试中人人都能得100分，那么这次考试是没有意义的；第二个问题是很多"新"数据集就是用已有的数据集整合整合做二次开发，就那么多源题目折腾来折腾去，导致训练集与测试基准重合（污染）。为此，一个高质量的、有难度的、新题目的数据集就显得尤为重要。

数据来源



Math StachExchange负责提供高难度题目，
NuminaMath-CoT负责提供简单题目

示例数据

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

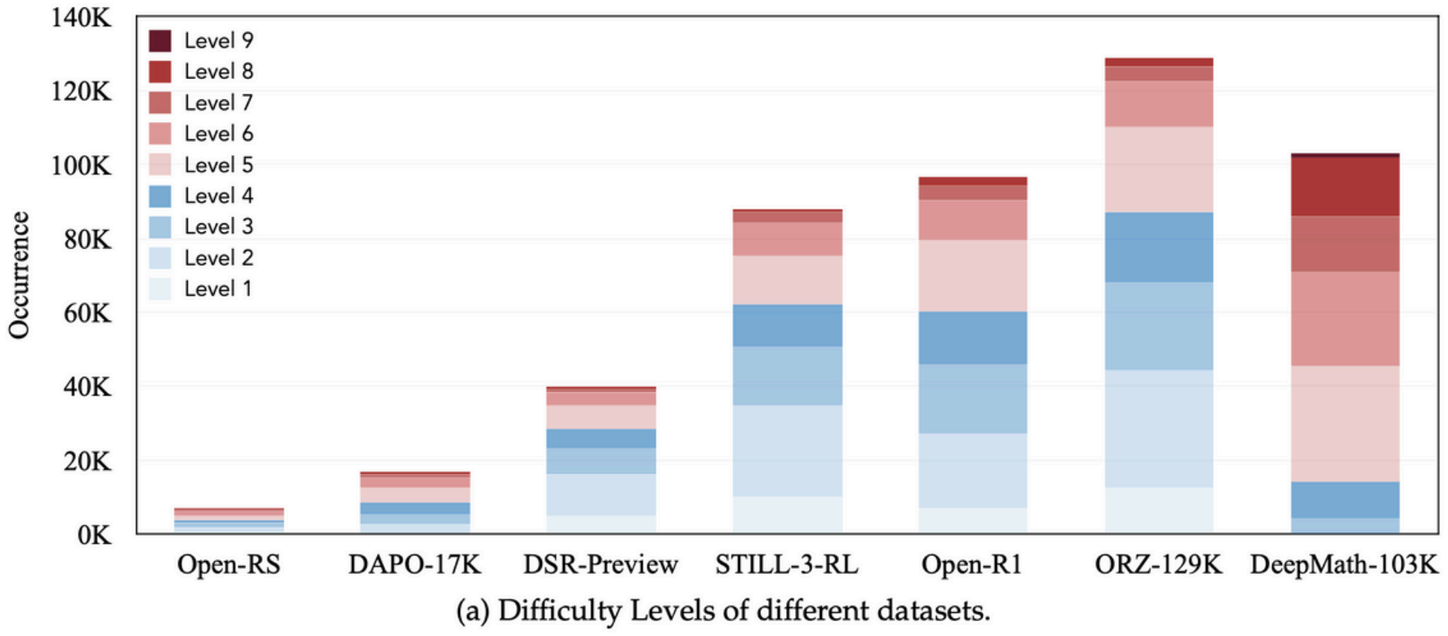
R1 Solution 2: Okay, so I need to calculate the line integral …. Hmm, first things first, let me recall what line integrals are about … Thus, the value of the line integral is: 2π

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

Figure 2: A data sample from DeepMath-103K.

题目主题

题目难度对比



(a) Difficulty Levels of different datasets.

难度定义来自AoPS网站，然后用GPT-4o对数据集中题目做few-shot 标注难度

部分实验结果

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.
↳ Open-Reasoner-Zero-7B (Hu et al., 2025)	81.8	58.9	47.9	38.4	15.6	14.4	40.
↳ Qwen-2.5-7B-SRL-Zoo (Zeng et al., 2025a)	77.0	55.8	41.0	41.2	15.6	8.7	33.
↳ DeepMath-Zero-7B (Ours)	**85.5**	**64.7**	**51.0**	**45.3**	**20.4**	**17.5**	**42.**
Qwen-2.5-Math-7B (Team, 2024)	46.9	31.9	15.8	15.5	11.2	4.4	22.
↳ Qwen-2.5-Math-7B-SRL-Zoo (Hu et al., 2025)	75.8	59.7	37.4	29.9	24.0	10.2	36.
↳ Qat-Zero-7B (Liu et al., 2025)	80.0	66.7	43.4	40.8	32.7	11.7	40.
↳ Eurus-2-7B-PRIME (Cui et al., 2025)	80.2	64.7	44.9	42.1	19.0	12.7	38.
↳ DeepMath-Zero-Math-7B (Ours)	**86.9**	**74.7**	**52.3**	**49.5**	**34.2**	**23.5**	**46.**
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.
↳ DeepScaleR-1.5B-Preview (Luo et al., 2025)	89.4	80.3	60.9	42.2	**42.3**	29.6	**46.**
↳ Still-3-1.5B-Preview (Chen et al., 2025)	86.6	75.8	55.7	38.7	30.8	24.6	43.
↳ DeepMath-1.5B (Ours)	**89.9**	**82.3**	**61.8**	**42.5**	37.3	**30.8**	46.
OpenMath-Nemotron-1.5B (Moshkov et al., 2025)	91.8	90.5	70.3	26.3	61.3	50.6	56.
↳ DeepMath-Omn-1.5B (Ours)	**93.2**	**94.2**	**73.4**	**28.3**	**64.0**	**57.3**	**58.**

用DeepMath-103K训练的模型效果着实不错，佐证了数据集的质量。