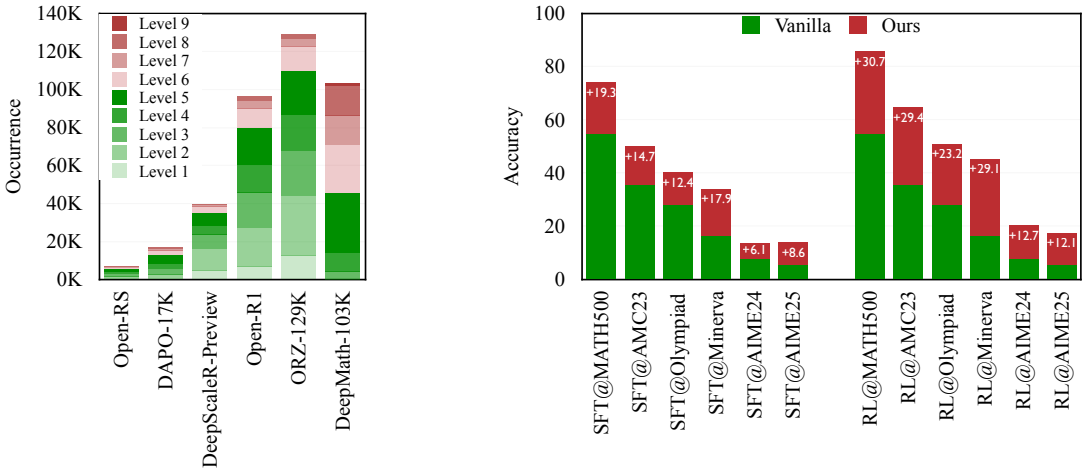
DeepMath-103K: A Large-Scale, Challenging, Decontaminated, and Verifiable Mathematical Dataset for Advancing Reasoning

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] , **and Dong Yu**[1]

[1]Tencent [2]Shanghai Jiao Tong University

https://github.com/zwhe99/DeepMath



(a) Difficulty Levels (b) Improvement of SFT and RL using DeepMath-103K

Figure 1: We introduce **DeepMath-103K**, a large-scale dataset of challenging mathematical problems. DeepMath-103K is more challenging compared to existing datasets (Figure a). All problems have both verified final answer and multiple solution paths, supporting a broad spectrum of training paradigms that substantially improves reasoning performance (Figure b).

Abstract

The capacity for complex mathematical reasoning is a key benchmark for artificial intelligence. While reinforcement learning (RL) applied to LLMs shows promise, progress is significantly hindered by the lack of large-scale training data that is sufficiently challenging, possesses verifiable answer formats suitable for RL, and is free from contamination with evaluation benchmarks. To address these limitations, we introduce DeepMath-103K, a new, large-scale dataset comprising approximately 103K mathematical problems, specifically designed to train advanced reasoning models via RL. DeepMath-103K is curated through a rigorous pipeline involving source analysis, stringent decontamination against numerous benchmarks, and filtering for high difficulty (primarily Levels 5-9), significantly exceeding existing open resources in challenge. Each problem includes a verifiable final answer, enabling rule-based RL, and three distinct R1-generated solutions suitable for diverse training paradigms like supervised fine-tuning or distillation. Spanning a wide range of mathematical topics, DeepMath-103K promotes the development of generalizable reasoning. We demonstrate that models trained on DeepMath-103K achieve significant improvements on challenging mathematical benchmarks, validating its effectiveness. We release DeepMath-103K publicly to facilitate community progress in building more capable AI reasoning systems.

[*]Equal Contribution. The work was done when Zhiwei, Xingyu, and Yue were interning at Tencent.
[†]Correspondence to: Zhaopeng Tu <zptu@tencent.com> and Rui Wang <wangrui12@sjtu.edu.cn>.

1 Introduction

Mathematical reasoning stands as a cornerstone capability for advanced artificial intelligence, serving as a critical proving ground for models aiming to emulate sophisticated human problem-solving (Kojima et al., 2022; Wei et al., 2022). Recent strides, particularly leveraging reinforcement learning (RL) with large language models (LLMs), have demonstrated significant potential in tackling complex mathematical problems demanding logical deduction, symbolic manipulation, and multi-step reasoning (Jaech et al., 2024; Guo et al., 2025; Team, 2024; xAI, 2025; Google, 2025). Notably, methods like RL-Zero (Guo et al., 2025), which employ online RL guided by binary rewards from verifiable answers, have surpassed traditional supervised fine-tuning approaches.

Despite this promise, the advancement of these powerful RL techniques is critically hampered by a bottleneck: the scarcity of suitable training data. Existing mathematical datasets (Hendrycks et al., 2021b; Cobbe et al., 2021a; Yu et al., 2024; Toshniwal et al., 2024) often fall short in several key aspects. They may lack the **extreme difficulty** needed to push the boundaries of current models, miss the **verifiable answer format** essential for rule-based RL reward schemes, suffer from **contamination** with standard evaluation benchmarks (compromising evaluation integrity), or are simply insufficient in **scale**, particularly concerning highly challenging problems. While human-annotated datasets tailored for RL (Wang et al., 2024; Face, 2025; Hu et al., 2025; Luo et al., 2025b; Yu et al., 2025; Dang & Ngo, 2025; Albalak et al., 2025) provide valuable insights, they often struggle with scale and capturing the extreme difficulty characteristic of competitive mathematics, which is necessary for training state-of-the-art reasoners.

To bridge this critical gap, we introduce **DeepMath-103K**, a novel, large-scale mathematical dataset specifically engineered to accelerate the development of advanced reasoning models via reinforcement learning. DeepMath-103K is distinguished by its high concentration of **challenging mathematical problems**, significantly surpassing the difficulty distribution prevalent in existing open datasets (Figure 1a). Our dataset construction methodology incorporates rigorous **data decontamination** against a comprehensive suite of benchmarks (Figure 4) to ensure trustworthy evaluation, alongside filtering for problems predominantly at high difficulty levels (≥ 5).

Crucially, every problem within DeepMath-103K features a **verifiable final answer**, directly enabling the application of rule-based reward functions in RL frameworks. Furthermore, each problem is accompanied by **three distinct R1-generated solutions** (Guo et al., 2025), offering rich data for diverse training paradigms such as supervised fine-tuning, reward modeling, or model distillation (Figure 2). The dataset covers a **broad spectrum of mathematical topics** (Figure 3), ranging from foundational concepts to advanced subjects, thereby promoting the development of generalizable reasoning abilities. Comprising approximately 103K problems — including 95K curated challenging examples (Levels 5-10) and 8K supplementary problems (Levels 3-5) — DeepMath-103K represents a significant contribution toward equipping the community with the resources required to train highly proficient mathematical reasoners.

We validate the effectiveness of DeepMath-103K by demonstrating that models trained on it achieve substantial performance gains on several demanding mathematical reasoning benchmarks. This work furnishes a vital, openly accessible resource, tackling the urgent need for large-scale, challenging, verifiable, and clean mathematical data essential for propelling the next generation of AI reasoning systems.

Our main contributions are as follows:

- We design and release DeepMath-103K, a novel large-scale mathematical dataset specifically curated for training advanced reasoning models via reinforcement learning, characterized by its high difficulty, verifiable answers, multiple diverse solutions per problem, and rigorous decontamination.
- We detail a meticulous data curation pipeline, encompassing source analysis, extensive decontamination against standard benchmarks for evaluation integrity, difficulty filtering, and robust answer verification.

- We demonstrate the effectiveness of DeepMath-103K by showing that models trained on it achieve significant improvements on multiple challenging mathematical reasoning benchmarks, particularly using RL-Zero techniques enabled by the dataset's structure.

2 Overview of 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

Figure 2: An example data sample from the DeepMath-103K dataset, illustrating its components.

Each data sample in DeepMath-103K is intentionally structured to be comprehensive, supporting a variety of downstream applications in mathematical reasoning research. As illustrated in Figure 2, a single sample includes the following components:

- *Question*: The mathematical problem statement.
- *Final Answer*: A verifiable final answer, crucial for enabling rule-based reward functions in RL settings.
- *Difficulty*: A numerical difficulty score, which facilitates techniques like difficulty-aware training (e.g., curriculum learning) or adaptive compute allocation based on problem complexity (Wang et al., 2025; Chen et al., 2024).
- *Topic*: A hierarchical topic classification for the problem, enabling topic-specific analysis or training.
- *R1 Solutions*: Three distinct reasoning paths generated by the DeepSeek-R1 model (Guo et al., 2025), suitable for diverse training paradigms.

DeepMath-103K possesses several key characteristics that make it particularly suitable for advancing mathematical reasoning research:

Higher Difficulty DeepMath-103K includes mathematical problems spanning difficulty levels 3 through 10. The dataset comprises 95K challenging problems (levels 5-10) specifically curated for this research, augmented with an additional 8K problems (levels 3-5) sourced from SimpleRL (Zeng et al., 2025b) to ensure broader difficulty coverage. For comparison, we analyzed and labeled the difficulty levels of several existing datasets commonly used for RL training in this domain: Open-RS (Dang & Ngo, 2025), DAPO-17K (Yu et al., 2025), DeepScaleR-Preview (Luo et al., 2025b), ORZ-129K (Hu et al., 2025), and Open-R1 (Face, 2025). Figure 1a illustrates the difficulty distributions across these datasets. As depicted, DeepMath-103K exhibits a significantly more challenging problem distribution, containing a substantially higher proportion of problems at difficulty levels 5 and above compared to the other benchmark datasets. This focus on higher difficulty is intended to push the reasoning limits of current models.

Broad Topical Diversity A key characteristic of DeepMath-103K, alongside its high difficulty, is its extensive topical diversity across the mathematical landscape. We systematically categorized each problem using a hierarchical topic structure, following the methodology established by Gao et al. (2024). As illustrated in Figure 3, this classification reveals that DeepMath-103K draws problems from a multitude of core mathematical areas. The dataset's scope ranges from fundamental topics such as Prealgebra and Plane Geometry to sophisticated domains like Abstract Algebra (including Group Theory and Field Theory) and advanced Calculus (covering Differential Equations and Applications of Integrals, among others). This broad topical foundation ensures that models trained on DeepMath-103K are exposed to a rich variety of mathematical concepts and problem-solving paradigms, thereby fostering the development of more robust and widely generalizable reasoning skills.

Figure 3: Hierarchical breakdown of covered mathematical topics in DeepMath-103K.

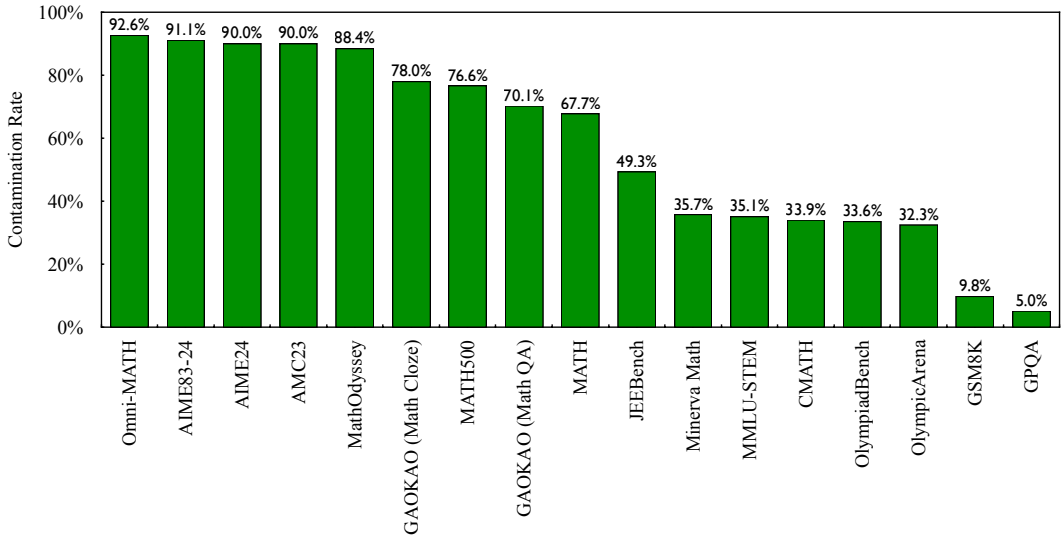


Figure 4: Contamination rates of common mathematical and STEM benchmarks detected in the raw data sources before decontamination.

Rigorous Data Decontamination DeepMath-103K was constructed exclusively using the training splits of existing open-source datasets, with careful avoidance of any known test set materials. However, our preliminary analysis revealed that these source datasets exhibit concerning levels of contamination with problems from commonly used evaluation benchmarks. As illustrated in Figure 4, the contamination rates (defined as the percentage of benchmark test samples found within the raw data pool aggregated from source training splits) were notably high: reaching 90% for AIME24 and AMC23, 76.6% for MATH500, 35.7% for Minerva Math, and 33.6% for OlympiadBench. Recognizing that these benchmarks are frequently employed for model evaluation, DeepMath-103K underwent a rigorous decontamination procedure. This process systematically identified and

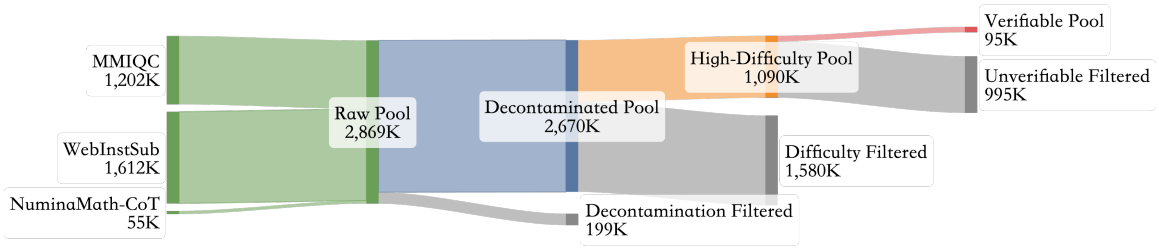


Figure 5: The data curation pipeline for DeepMath-103K. Starting with an initial pool of 2,869K raw questions, successive stages of data decontamination, difficulty filtering (retaining levels $\geq$5), and answer verifiability filtering yield 95K problems. These are then combined with 8K problems from SimpleRL (Zeng et al., 2025b) to form the final DeepMath-103K dataset.

removed problems that overlap with these standard evaluation sets, ensuring the integrity of future benchmark results obtained using models trained on DeepMath-103K.

Suitable for Diverse Training Paradigms A core advantage of DeepMath-103K is its comprehensive data structure with both a verified final answer and multiple solution paths, supporting a broad spectrum of training paradigms and research methodologies.

- *Supervised Fine-Tuning:* By providing three distinct *R1-generated solutions* for each problem, DeepMath-103K enables the creation of rich supervised training corpora. These solutions offer multiple valid approaches to the same question, allowing a model to learn diverse problem-solving strategies. In contrast to datasets that present only a single correct path, the variety in DeepMath-103K's solutions helps LLMs better generalize to unseen problems.
- *Model Distillation:* Advanced teacher–student paradigms often rely on having multiple labeled solution paths for individual questions. Through DeepMath-103K's three solution trajectories, a larger teacher model can effectively impart varied problem-solving styles to a smaller student model, enhancing the student's coverage of different reasoning heuristics and strategies.
- *Rule-based Reinforcement Learning (e.g., RL-Zero):* The availability of a verifiable final answer in each problem allows for straightforward reward assignment, as models can be directly evaluated on whether their predicted result matches the correct solution. This binary feedback, essential for RL-based methods such as RL-Zero (Guo et al., 2025), fosters deeper reasoning by encouraging improvements specifically targeted at correctness.
- *Reward Modeling:* With multiple valid solution paths, it becomes possible to design reward models that differentiate between high- and low-quality reasoning steps or compare alternative solution routes. This not only refines policy gradients in RL frameworks but also aids in re-ranking or scoring candidate solutions in a multi-step decoding pipeline (xAI, 2025).

Taken together, these features make DeepMath-103K exceptionally flexible for cutting-edge research on AI-driven mathematical reasoning, supporting both direct RL with correctness-based rewards and more nuanced learning frameworks that benefit from multiple, validated solution paths.

3 Construction of DeepMath-103K

This section details the meticulous data curation process used to construct DeepMath-103K, illustrated in Figure 5. The process comprises four primary stages:

1. **Source Analysis and Collection:** Identifying and collecting mathematically challenging problems by analyzing the difficulty distributions of existing open data sources.

2. **Data Decontamination:** Rigorously decontaminating the collected data to remove potential overlaps with standard evaluation benchmarks, ensuring evaluation integrity.

3. **Difficulty Filtering:** Filtering the decontaminated problems based on difficulty, retaining only those assessed at level 5 or higher to focus on challenging content.

4. **Answer Verification:** Ensuring each curated problem possesses a verifiable final answer, consistently validated across multiple solution paths generated by DeepSeek-R1.

Overall, this curation pipeline ensures that DeepMath-103K is largely free from benchmark contamination and concentrates on challenging mathematical problems suitable for advanced reasoning model training. The entire procedure involved significant computational resources, requiring an expenditure of **138,000 US dollars in GPT-4o API fees** and a total of **127,000 H20 GPU hours**.

Stage 1: Source Analysis and Collection. To identify data sources rich in challenging problems, we began by analyzing the landscape of existing open mathematical reasoning datasets. These datasets utilize diverse collection methodologies. For instance, datasets such as MetaMathQA (Yu et al., 2024), dart-math-hard (Tong et al., 2024), and OpenMathInstruct-2 (Toshniwal et al., 2024) primarily focus on augmenting problems and solutions derived from established datasets like GSM8K (Cobbe et al., 2021a) and MATH (Hendrycks et al., 2021b). In contrast, datasets like NuminaMath-CoT (LI et al., 2024), MMIQC (Liu et al., 2024), and WebInstructSub (Yue et al., 2024) source content more broadly from the web, gathering materials such as exercises and discussions from online platforms (e.g., Math Stack Exchange).

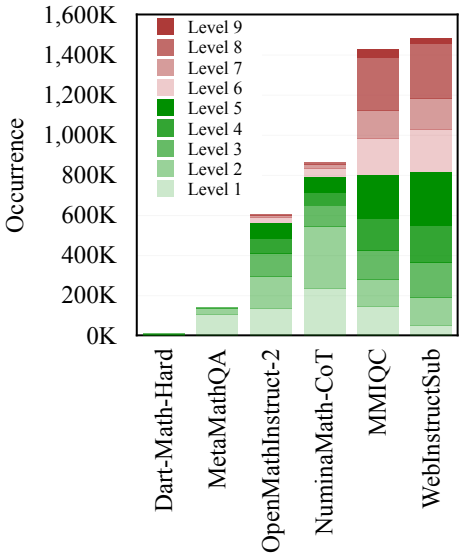


Figure 6: Difficulty distributions of various open mathematical reasoning datasets considered as potential sources.

We follow Gao et al. (2024) to estimate the difficulty distribution of these potential source datasets, as shown in Figure 6. This analysis revealed distinct patterns: datasets derived mainly from GSM8K and MATH (MetaMathQA, dart-math-hard, OpenMathInstruct-2), along with NuminaMath-CoT, exhibited distributions heavily skewed towards lower difficulty levels (levels 1-5). Conversely, datasets sourced more broadly from web content, specifically MMIQC and WebInstructSub, displayed significantly flatter distributions with a larger proportion of problems in the mid-to-high difficulty range (levels 5-9). Based on this finding, we selected MMIQC and WebInstructSub as our primary data sources due to their higher concentration of challenging problems. We also included NuminaMath-CoT to enhance the topical diversity of the initial collection. After applying basic filtering, this selection process yielded a raw pool of 2,869K questions.

Stage 2: Data Decontamination. As indicated by the high contamination rates observed in common benchmarks (Figure 4), a rigorous data decontamination process was crucial for ensuring the integrity of DeepMath-103K. We performed decontamination against a comprehensive suite of mathematics and STEM benchmarks, including MATH (Hendrycks et al., 2021b), AIME (MAA, a), AMC (MAA, b), Minerva Math (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), Omni-MATH (Gao et al., 2024), MathOdyssey (Fang et al., 2024), GAOKAO (Zhong et al., 2023), JEEBench (Arora et al., 2023), MMLU-STEM (Hendrycks et al., 2021a), CMATH (Wei et al., 2023), OlympicArena (Huang et al., 2024), GSM8K (Cobbe et al., 2021a), and GPQA (Rein et al., 2024). We adopted the decontamination method proposed by Toshniwal et al. (2024) to effectively remove potential paraphrases of benchmark questions. This involved the following steps:

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?
MATH500	Sorry to bother you today, but I came across this question: There are finitely many primes p for which the congruence $8x \equiv 1 \pmod{p}$ has no solutions x. Determine the sum of all such p. I first thought the answer was 0, since they didn't say that x had to be an integer, but apparently, it did. I don't know how to proceed from here, any solutions? Thanks for taking the time to read this!	There are finitely many primes p for which the congruence $8x \equiv 1 \pmod{p}$ has no solutions x. Determine the sum of all such p.

Table 1: Examples of contamination detected between the raw data pool and benchmarks using semantic comparison. Colors highlight conceptual or textual similarities leading to flagging.

1. For each candidate question in our raw dataset, we employed embedding similarity search (using paraphrase-multilingual-MiniLM-L12-v2 (Reimers & Gurevych, 2019)) to identify the top-k ($k = 5$) most similar examples from the aggregated test sets of all targeted benchmarks.
2. Each candidate question was then compared against its top-k retrieved benchmark examples using an LLM-Judge (Llama-3.3-70B-Instruct (Grattafiori et al., 2024)) to determine if they constituted identical questions or paraphrases. To mitigate positional bias, we evaluated each pair twice, swapping the order of the questions, resulting in $2k$ comparisons per candidate. If any of these $2k$ comparisons indicated a potential paraphrase or duplicate, the candidate question was discarded.

This semantic approach aims to identify not only exact duplicates but also near-duplicates and paraphrased questions that might otherwise overlap with evaluation sets.

Table 1 illustrates the effectiveness of this semantic decontamination compared to simple lexical matching. For instance, the AIME24 and AMC23 examples show that questions conveying similar mathematical concepts were correctly flagged as contaminated, despite differences in numerical values and phrasing. The MATH500 example demonstrates the detection of a near-identical question embedded within conversational text. These cases underscore the capability of our method to identify subtle conceptual overlaps, thereby yielding a dataset largely free from leakage relative to common evaluation benchmarks.

Stage 3: Difficulty Filtering. Prior work Zeng et al. (2025a) highlights the importance of aligning RL training data difficulty with the target model's reasoning capabilities, noting that powerful models benefit significantly from exposure to highly challenging problems. Building on this insight, our curation process for DeepMath-103K focuses on selecting problems that represent a significant reasoning challenge. To quantify difficulty, we adopted the approach detailed in Omni-MATH (Gao et al., 2024). We assigned a difficulty level to each decontaminated problem by prompting GPT-4o

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 $n \geq k$ sides, as the angle of rotation θ varies from 0 to 2π. Provide the sequence $s(k, n)$ for all $n \geq k$.
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 $X \neq \emptyset$ 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 $\varphi(l) \in M$.

Table 2: Examples of geometry problems retained by the difficulty filtering process (Level ≥ 5).

based on the annotation guidelines provided by the Art of Problem Solving (AoPS)[1]. 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. Subsequently, we applied a strict filtering criterion, retaining only those problems with an estimated difficulty level of 5 or higher. Table 2 showcases examples of geometry problems that passed this filtering stage, illustrating how increasing difficulty levels often correlate with greater conceptual depth and reasoning complexity.

Stage 4: Answer Verification. The availability of verifiable final answers is crucial for enabling rule-based reward mechanisms in RL, which helps mitigate reward hacking and has been instrumental in training successful reasoning models like DeepSeek-R1 (Guo et al., 2025). However, reliably constructing such answers presents two primary challenges:

1. Some question types, such as mathematical proofs or open-ended explorations, inherently lack a unique, easily verifiable final result.
2. Certain answers are excessively complex (e.g., lengthy expressions or intricate notation), making them challenging or even infeasible for automated rule-based verification.

To address these issues, we implemented a rigorous two-stage verification process:

1. **Question Formatting and Filtering:** We utilized GPT-4o to process the raw questions. Problem types inherently unsuitable for verification (e.g., proofs) were discarded. Questions phrased conversationally were automatically rewritten into a standardized format seeking a single, specific numerical or symbolic answer.
2. **Answer Verification via Consistency Check:** For questions successfully passing the above step, we generated **three distinct solution paths using the DeepSeek-R1 model**. 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). We enforced strict consistency: only problems where all extracted final answers were identical were retained in the final dataset.

[1] https://artofproblemsolving.com/wiki/index.php/AoPS_Wiki:Competition_ratings

Model	MATH500	AMC23	Olympiad Math	Minerva Bench	AIME24	AIME25
Qwen2.5-Math-7B-Instruct	83.2	59.2	42.6	41.7	12.1	11.0
Qwen2.5-7B-Base	54.8	35.3	27.8	16.2	7.7	5.4
Supervised Fine-Tuning						
1 R1 Solution	69.2	47.3	35.9	29.8	12.3	8.7
3 R1 Solutions	74.1	50.0	40.2	34.1	13.8	14.0
Zero Reinforcement Learning						
Open-Reasoner-Zero-7B	81.8	58.9	47.9	38.4	15.6	14.4
Qwen-2.5-7B-SimpleRL-Zoo	77.0	55.8	41.0	41.2	15.6	8.7
DeepMath-Zero-7B (Ours)	**85.5**	**64.7**	**51.0**	**45.3**	**20.4**	**17.5**
R1-Distill-Qwen-1.5B (Instruct)	84.7	72.0	53.1	36.6	29.4	24.8
Reinforcement Learning						
DeepScaleR-1.5B-Preview	**89.4**	80.3	**60.9**	**42.2**	**42.3**	29.6
Still-3-1.5B-Preview	86.6	75.8	55.7	38.7	30.8	24.6
DeepMath-1.5B (Ours)	89.0	**81.6**	60.1	40.6	39.8	**30.8**

Table 3: Results of RL and SFT using different training datasets. "DeepMath" denotes models trained using DeepMath-103K. Performance is evaluated using pass@1 (n=16) accuracy. We also list the results of Qwen2.5-Math-7B-Instruct for reference.

This combined approach of question standardization and multi-solution answer consistency checking ensures that every problem included in DeepMath-103K possesses a final answer that is robustly verifiable using automated rules.

4 Effectiveness of DeepMath-103K

4.1 Results on Challenging Mathematical Benchmarks

To empirically validate the effectiveness of DeepMath-103K for training advanced mathematical reasoning models, we conducted experiments utilizing different training paradigms. We trained models starting from two distinct initial checkpoints: a base LLM Qwen-2.5-7B-Base and a model already supervised fine-tuned for mathematical reasoning – R1-Distill-Qwen-1.5B. We evaluated performance across a suite of challenging mathematical benchmarks: MATH500 (Hendrycks et al., 2021b), AIME 2024-2025 (MAA, a), AMC 2023 (MAA, b), Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). The results summarized in Table 3 clearly demonstrate the benefits of training with DeepMath-103K.

Supervised fine-tuning on DeepMath-103K significantly enhances base model performance, with multiple solutions yielding further gains. As detailed in Section 2, each problem in DeepMath-103K includes three distinct R1-generated solutions, facilitating supervised fine-tuning (SFT). We fine-tuned Qwen-2.5-7B-Base using either the first R1 solution or all three solutions provided in DeepMath-103K. As shown in Table 3, SFT yields substantial improvements over the base model across all benchmarks. For instance, using just one R1 solution boosts MATH500 accuracy from 54.8% to 69.2%. Leveraging all three distinct solutions further improves performance (e.g., 74.1% on MATH500, 14.0% on AIME25), demonstrating the value of exposing the model to diverse problem-solving strategies.

DeepMath-103K enables state-of-the-art performance for the base model via RL-Zero. The RL-Zero (Guo et al., 2025) trains models using online RL with binary rewards derived directly from verifiable final answers. This approach aligns perfectly with DeepMath-103K's core design principle of providing a rigorously verified final answer for every problem (Stage 4 in Section 3), enabling direct application of rule-based rewards crucial for advancing mathematical reasoning. We trained

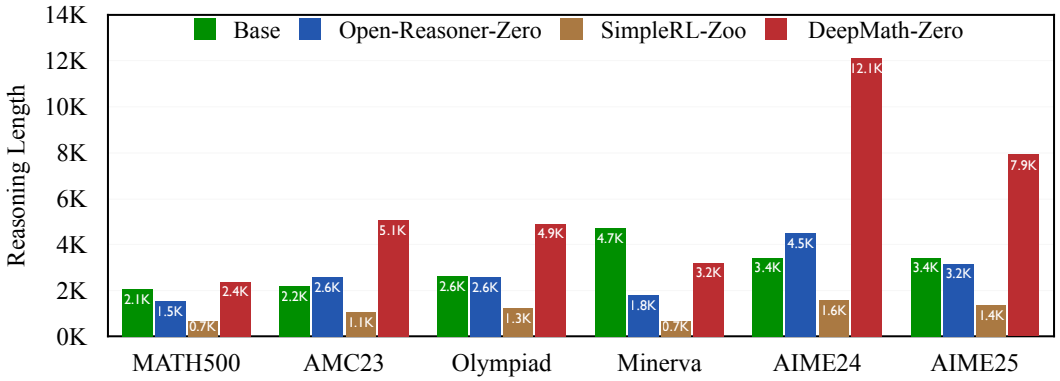


Figure 7: Average reasoning length (number of tokens) on different benchmarks for models trained with RL-Zero using various datasets, starting from Qwen-2.5-7B-Base.

the Qwen2.5-7B-Base model using RL-Zero on DeepMath-103K (denoted "DeepMath-Zero"). For comparison, we included results from models trained using RL-Zero on other public RL datasets: Open-Reasoner-Zero-7B (Hu et al., 2025) and Qwen-2.5-7B-SimpleRL-Zoo (Zeng et al., 2025a) for the 7B model. Table 3 shows that DeepMath-Zero-7B consistently and significantly outperforms both the SFT models and models trained on other RL datasets across all benchmarks.

In addition, we also conducted experiments on R1-Distill-Qwen-1.5B, a model already supervised fine-tuned for mathematical reasoning. For comparison, we included results of DeepScaleR-1.5B-Preview (Luo et al., 2025b) and Still-3-1.5B-Preview (Face, 2025) for the 1.5B model. As shown in Table 3, DeepMath-Zero-1.5B achieves state-of-the-art results among the compared 1.5B models on AMC23 (81.6%) and AIME25 (30.8%), and performs competitively at the top on the remaining benchmarks.

The significant performance gains validate DeepMath-103K's design principles for advancing mathematical reasoning. These results strongly demonstrate DeepMath-103K's effectiveness. The substantial improvements, particularly with RL-Zero, underscore the value of DeepMath-103K's core characteristics. The dataset's focus on high-difficulty problems pushes models towards more sophisticated reasoning. The availability of verifiable final answers enables effective optimization via rule-based RL. Furthermore, the rigorous decontamination process ensures that the observed gains reflect genuine reasoning improvements, not benchmark leakage. The scale and topical diversity likely contribute to the robustness observed across various benchmarks. Overall, these experiments confirm that DeepMath-103K serves as a powerful resource for training highly capable mathematical reasoning models.

4.2 Analysis of RL-Zero Using DeepMath-103K

In this section, we present a qualitative analysis to provide insights into how training on DeepMath-103K with RL-Zero enhances mathematical reasoning capabilities, focusing on DeepMath-Zero-7B.

Training on DeepMath-103K often encourages models to generate substantially longer and more detailed reasoning steps, particularly on highly complex benchmarks. We analyzed the average length (in tokens) of the solutions generated by DeepMath-Zero-7B compared to the base model and models trained on other RL datasets (Figure 7). DeepMath-Zero-7B produces markedly longer reasoning chains on several benchmarks, especially the more challenging ones like AIME24 (12106 tokens vs. 4503 for Open-Reasoner-Zero and 1582 for Qwen-2.5-7B-SimpleRL-Zoo), AIME25 (7920 vs. 3177 and 1365). While solution length is not a direct measure of quality, this significant increase suggests that training on DeepMath-103K's challenging problems prompts the model to engage in more elaborate, step-by-step derivations, potentially reflecting deeper processing required to

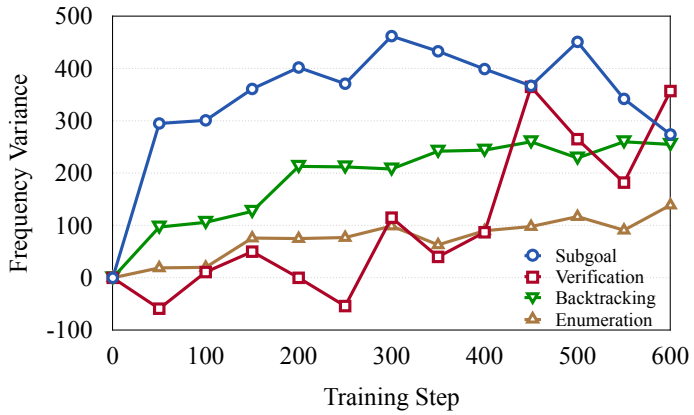


Figure 8: Variance (change from initial state) of beneficial cognitive behaviors observed during RL-Zero training on DeepMath-103K.

solve complex tasks. SimpleRL-Zoo, focusing on simpler problems, leads to the shortest solutions, aligning with expectations.

Models trained on DeepMath-103K exhibit increased use of beneficial cognitive behaviors during RL-Zero training. We followed the methodology of Gandhi et al. (2025); Zeng et al. (2025a) to track the emergence of four key cognitive behaviors associated with effective problem-solving during the RL-Zero training of DeepMath-Zero-7B:

- *Subgoal Setting:* Breaking down complex problems into smaller, manageable steps (e.g., "To solve this, we first need to...").
- *Verification:* Systematically checking intermediate results or reasoning steps (e.g., "Let's verify this result by...").
- *Backtracking:* Explicitly revising approaches upon identifying errors or dead ends (e.g., "This approach won't work because... Let's try...").
- *Enumeration:* Solving problems by exhaustively considering multiple cases or possibilities.

Figure 8 shows the change in the frequency of these behaviors compared to the initial model over the course of training. We observe notable increases in all beneficial cognitive behaviors as training progresses. This suggests that optimizing on DeepMath-103K's challenging and verifiable problems encourages the model to adopt more structured and robust problem-solving strategies, moving beyond simple pattern matching towards more deliberate reasoning processes involving planning, checking, and adapting its approach. The emergence of these behaviors aligns with the goal of developing more general and capable mathematical reasoners.

5 Related Work

LLM Complex Reasoning LLMs use Chain-of-Thought (CoT) prompting (Wei et al., 2022) to generate step-by-step reasoning, significantly improving their effectiveness in reasoning-intensive tasks. A line of strategies tends to directly improve the LLM CoT performance on reasoning tasks through demonstrations (Zhou et al., 2022; Zhang et al., 2023; Brown et al., 2020; Wang et al., 2022), while other works contributes to the CoT structures (Yao et al., 2023; Chen et al., 2023), planning (Wang et al., 2023), and difficulty (Fu et al., 2023).

Recent advancement in inference-time scaling (Snell et al., 2024; Brown et al., 2024; Wu et al., 2024) significantly boosts the performance of LLM reasoning on complex tasks. Building on this, online reinforcement learning (RL) with correctness rewards has emerged as a promising direction for

reasoning models (Guo et al., 2025). By training models to optimize for correct final answers, these methods yield state-of-the-art performance on complex reasoning tasks (Jaech et al., 2024; Guo et al., 2025; Team, 2024; xAI, 2025; Google, 2025). Motivated by this, we propose DeepMath-103K dataset of large-scale and verifiable answers, which is suitable for RL of LLM complex reasoning capabilities.

Mathematical Reasoning Datasets and Beyond A typical paradigm to enhance the LLM reasoning capabilities is to conduct supervised fine-tuning on human-annotated reasoning data. Those mathematical reasoning data of question-answer pairs are mainly mined from the web corpus (LI et al., 2024; Zeng et al., 2024; Li et al., 2024; Yu et al., 2024; Yue et al., 2024). Correspondingly, with improved reasoning capabilities of LLMs, the evaluation benchmark datasets focus on assessing LLMs on problem-solving accuracy, step-by-step reasoning, and robustness on more challenging and difficulty levels beyond the human experts (Hendrycks et al., 2021b; Lewkowycz et al., 2022; He et al., 2024; Gao et al., 2024; Fang et al., 2024; Zhong et al., 2023; Arora et al., 2023; Hendrycks et al., 2021a; Wei et al., 2023; Huang et al., 2024; Cobbe et al., 2021a; Rein et al., 2024; Cobbe et al., 2021b; MAA, a;b; Glazer et al., 2024).

In the RL-Zero era, the scarcity of **challenging, diverse, and verifiable** reasoning questions impedes reinforcement learning-based zero-shot LLM reasoning. To address this, we introduce DeepMath-103K, a novel dataset of carefully curated questions spanning varied difficulty levels and topics. Additionally, DeepMath-103K is designed to be contamination-free with respect to evaluation benchmarks, mitigating concerns about test set leakage.

RL-Zero Reinforcement Learning RL-Zero (Guo et al., 2025) is a streamlined framework for developing reinforcement learning (RL) capabilities in base LLMs from scratch. It elucidates LLM reasoning by employing RL principles, using interactions with the environment, i.e., the correctness of reasoning on question examples, to bolster exploration and enhance reasoning performance. Recent progress in large language model (LLM) reasoning has been profoundly shaped by reinforcement learning (RL) techniques (Jaech et al., 2024; Guo et al., 2025; Team, 2024; xAI, 2025; Google, 2025).

Despite the successes of RL-Zero in advancing LLM reasoning, models tuned within this framework face challenges in reasoning efficiency. As observed by Chen et al. (2024) and Wang et al. (2025), these models tend to overthink simple questions and generate redundant CoTs through excessive verification, while underthinking complex ones leading to inconsistent performance across diverse problem difficulties. Based on the above analysis, recent studies focus on developing efficient RL-zero algorithms to enhance the LLMs' efficient reasoning capabilities (Chen et al., 2024; Team et al., 2025; Wang et al., 2025; Arora & Zanette, 2025; Yeo et al., 2025; Luo et al., 2025a). A core principle in these efforts is calibrating the LLM's response length to the actual difficulty of the task, which motivates us to construct DeepMath-103K dataset with difficulty annotation.

6 Conclusion

In this work, we addressed the critical bottleneck hindering progress in AI mathematical reasoning, particularly for reinforcement learning approaches: the lack of large-scale, sufficiently challenging, verifiable, and clean training data. We introduced DeepMath-103K, a novel dataset meticulously constructed to overcome these limitations. DeepMath-103K offers a substantial volume (103K problems) of primarily high-difficulty mathematical problems (Levels 5-9), each equipped with a verifiable final answer essential for rule-based RL, and further enriched with three distinct R1-generated solutions to support diverse training methodologies. Our rigorous curation process, featuring comprehensive decontamination against standard benchmarks, ensures the dataset's integrity for reliable model evaluation. We demonstrated the practical value of DeepMath-103K, showing that models trained on it exhibit significant performance gains on demanding mathematical reasoning tasks. By publicly releasing DeepMath-103K, we provide a vital resource to the research community, aimed at accelerating the development of more powerful and robust AI systems capable of tackling complex mathematical challenges.

References

Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models, 2025. URL https://arxiv.org/abs/2502.17387.

Daman Arora and Andrea Zanette. Training language models to reason efficiently. *arXiv preprint arXiv:2502.04463*, 2025.

Daman Arora, Himanshu Singh, and Mausam. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7527–7543, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.468. URL https://aclanthology.org/2023.emnlp-main.468.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=YfZ4ZPt8zd.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do not think that much for 2+3=? on the overthinking of o1-like llms, 2024. URL https://arxiv.org/abs/2412.21187.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021a.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021b. URL https://arxiv.org/abs/2110.14168.

Quy-Anh Dang and Chris Ngo. Reinforcement learning for reasoning in small llms: What works and what doesn't, 2025. URL https://arxiv.org/abs/2503.16219.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *arXiv preprint arXiv:2406.18321*, 2024.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yf1icZHC-l9.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-math: A universal olympiad level mathematic benchmark for large language models, 2024. URL https://arxiv.org/abs/2410.07985.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024. URL https://arxiv.org/abs/2411.04872.

Google. Gemini 2.0 flash thinking, 2025. URL https://cloud.google.com/vertex-ai/generative-ai/docs/thinking. Accessed on March 25, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model. https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero, 2025.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *arXiv preprint arXiv:2406.12753*, 2024. URL https://arxiv.org/abs/2406.12753.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with

language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf.

Chengpeng Li, Zheng Yuan, Hongyi Yuan, Guanting Dong, Keming Lu, Jiancan Wu, Chuanqi Tan, Xiang Wang, and Chang Zhou. Mugglemath: Assessing the impact of query and response augmentation on math reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10230–10258, 2024.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

Haoxiong Liu, Yifan Zhang, Yifan Luo, and Andrew Chi-Chih Yao. Augmenting math word problems via iterative question composing, 2024. URL https://arxiv.org/abs/2401.09003.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning, 2025a. URL https://arxiv.org/abs/2501.12570.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2, 2025b. Notion Blog.

MAA. American invitational mathematics examination (AIME). Mathematics Competition Series, a. URL https://maa.org/math-competitions/aime.

MAA. American mathematics competitions (AMC 10/12). Mathematics Competition Series, b. URL https://maa.org/math-competitions/amc.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, Nov 2024. URL https://qwenlm.github.io/blog/qwq-32b-preview/.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. DART-math: Difficulty-aware rejection tuning for mathematical problem-solving. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=zLU21oQjD5.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2609–2634, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.147. URL https://aclanthology.org/2023.acl-long.147/.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747*, 2022.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Thoughts are all over the place: On the underthinking of o1-like llms, 2025. URL https://arxiv.org/abs/2501.18585.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. Cmath: Can your language model pass chinese elementary school math test?, 2023.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*, 2024.

xAI. Grok: Artificial intelligence assistant, 2025. URL https://x.ai. Developed by xAI, accessed on March 25, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=N8N0hgNDRt.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Xiang Yue, Tianyu Zheng, Ge Zhang, and Wenhu Chen. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*, 37:90629–90660, 2024.

Liang Zeng, Liangjun Zhong, Liang Zhao, Tianwen Wei, Liu Yang, Jujie He, Cheng Cheng, Rui Hu, Yang Liu, Shuicheng Yan, et al. Skywork-math: Data scaling laws for mathematical reasoning in large language models–the story goes on. *arXiv preprint arXiv:2407.08348*, 2024.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025a. URL https://arxiv.org/abs/2503.18892.

Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient. https://hkust-nlp.notion.site/simplerl-reason, 2025b. Notion Blog.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=5NTt8GFjUHkr.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.