HighMATH: Evaluating Math Reasoning of Large Language Models in Breadth and Depth

Yan Liu[1], Minghui Zhang[1], Bojian Xiong[1], Yifan Xiao[1],
Yinong Sun[1], Yating Mei[1], Longyu Zeng[1], Jingchao Yang [1],
Yang Wang[1], Deyi Xiong[1,2]*,

[1] TJUNLP Lab, College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] University International College, Macau University of Science and Technology, Macau, China
{yan_liu, dyxiong}@tju.edu.cn

Abstract

With the rapid development of large language models (LLMs) in math reasoning, the accuracy of models on existing math benchmarks has gradually approached 90% or even higher. More challenging math benchmarks are hence urgently in need to satisfy the increasing evaluation demands. To bridge this gap, we propose HighMATH. Problems in HighMATH are collected according to 3 criteria: problem complexity, knowledge domain diversity and fine-grained annotations. We collect 5,293 problems from Chinese senior high school mathematics exams published in 2024, covering 8 subjects and 7 levels of difficulty, with each problem involving an average of more than 2.4 knowledge points. We conduct a thorough evaluation of latest LLMs on the curated HighMATH, including o1-like models. Evaluation results demonstrate that the accuracy of advanced LLMs on HighMATH is significantly lower than that on previous math reasoning benchmarks. This gap even exceeds 30%. Our results also suggest that properly trained smaller LLMs may have great potential in math reasoning. Our data is available at https://github.com/tjunlp-lab/HighMATH.

1 Introduction

LLMs have achieved significant progress in math reasoning (Li et al., 2024). When the challenging MATH (Hendrycks et al., 2021b) benchmark was initially proposed for evaluation, the accuracy of LLMs did not reach 20%. However, in just three years, LLMs are capable of achieving over 60% accuracy (Yang et al., 2024). Recently, with the emergence and application of techniques such as automatic process supervision (Wang et al., 2024), test-time scaling (Qi et al., 2024; Chen et al., 2024), and reinforcement learning (Rafailov et al., 2023; Ouyang et al., 2022), many models even achieve or exceed 90% accuracy on MATH (DeepSeek-AI et al., 2025). This rapid development of LLMs in math reasoning indicates that existing math benchmarks are no longer sufficient to evaluate and differentiate latest LLMs.

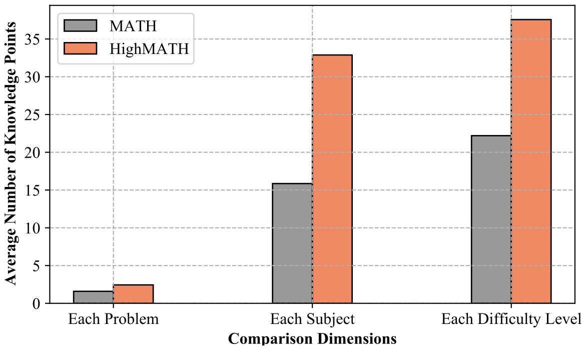

Figure 1: Comparison of HighMATH vs MATH. We count the number of knowledge points in each problem, subject, and difficulty level in the problems sampled from HighMATH and MATH. It can be seen that HighMATH contains a higher average number of knowledge points than MATH.

Particularly, commonly used benchmarks, such as GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), CMATH (Wei et al., 2023), and GAOKAO (Zhang et al., 2023), all suffer from these limitations. GSM8K, CMATH and FineMATH (Liu et al., 2024), which originate from elementary school math, present problems that are simple and do not allow for complex reasoning, making it difficult to deeply evaluate advanced LLMs. MATH, sourced from American high school math competitions, is quite challenging but typically presents problems with limited knowledge points, thus lacking the ability to evaluate LLMs in integrating multiple mathematical subjects. GAOKAO, C-EVAL (Huang et al., 2023), M3KE (Liu et al., 2023), and other comprehensive benchmarks (Hendrycks et al., 2021a; Zhong et al., 2024) are designed for multidisciplinary evaluation, where math is only one of these subjects, featuring

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10241–10253
November 4-9, 2025 ©2025 Association for Computational Linguistics

Benchmarks	**MATH**	**GAOKAO**	**C-EVAL**	**M3KE**	**Ours**
Language	En	Zh	Zh	Zh	Zh
Size	5,000	936	669	796	5,293
Problem Type	MWP	MCQ Fill-in-the-Blank MWP	MCQ	MCQ	MWP
Question Len.	191.13	185.92	76.28	46.24	158.63
Solution Len.	519.96	305.86	-	-	603.16
Subject/Level Label	✔	✗	✗	✗	✔

Table 1: Comparison of our benchmark against previous math benchmarks. MWP: Math Word Problem, MCQ: Multi-choice Questions. Question Len. and Solution Len. are the average character lengths we count.

fewer problems and coarser granularity.

To mitigate these challenges, we propose new Chinese math benchmark HighMATH. Our dataset offers a comprehensive and challenging assessment of LLMs in math reasoning by meticulously integrating a wide range of mathematical concepts, and knowledge into math problems with multiple reasoning steps. It contains 5,293 math problems, with each problem containing an average of more than 2.4 knowledge points, as shown in Figure 1. It covers all major areas of high school mathematics, including 8 domains, i.e., Function Derivatives, Counting Principles, Trigonometric Functions and Triangle Solutions, Plane Analytic Geometry, Sequences, Solid Geometry, Statistics and Probability, and others (Logic, Sets, Inequalities, Complex Numbers reasoning). In addition to subject annotation, we categorize problems into difficulty levels from 1 to 7. We also collect problems with multiple sub-questions into a subset called HighMATH-HARD. The average number of characters of problems and their solutions in HighMATH-HARD are 249 and 1,201 respectively, posing great challenges to LLMs in math reasoning.

We evaluate 14 LLMs (including both open- and closed-source models) on HighMATH. Compared to MATH, all models exhibit a significant drop in accuracy on HighMath, with even the latest reasoning language model, o1-mini, achieving only 52.53% accuracy. The o1-like model, DeepSeek-R1-distill-Qwen-32B, achieves only 31.22%. We also evaluate LLMs under the majority voting, pass@N settings, and with the assistance of a Python executor. However, these can only improve reasoning performance to a certain extent, indicating the difficulty of our benchmark.

Our contributions are summarized as follows.

- We propose a new challenging bechmark for evaluating LLMs in math reasoning. The dataset covers a wide range of mathematical concepts and typically examines multiple mathematical knowledge points in a single problem. Additionally, it has a sufficient number of problems, requires complex reasoning, and is finely annotated.

- We conduct thorough evaluations of the latest LLMs, including both o1-like models and those specifically trained for mathematical reasoning.

- We carry out an in-depth analysis of the evaluation results. The results show that the potential of small LLMs for mathematical reasoning is enormous. The pass@8 accuracy of the 1.5B model even surpasses that of many 7B models.

2 Related Work

Mathematical Reasoning Benchmarks. Our work is most closely related to MATH. We carefully analyze data in MATH (Hendrycks et al., 2021b) and find that, although it originates from American high school math competitions, it includes simple questions as well, such as those from the AMC10 competition, which contains ninth-grade mathematics. In terms of subject division, algebra content is categorized into prealgebra, intermediate algebra, and algebra, making up three of the seven categories. Additionally, almost all current LLMs use MATH training dataset to train their models. Previous analyses on contaminated samples discover that some existing training datasets, including the MATH training dataset, contain a significant number of problems that are highly similar in concept or structure to those in the test datasets (Yang et al., 2024). We speculate that these limitations may gradually render the MATH dataset inadequate for meeting the evaluation needs of LLMs. Detailed comparison of our benchmark to MATH is presented in Figure 1 and Table 1.

Datasets for evaluating mathematical reasoning in Chinese context are usually integrated into multidisciplinary or knowledge evaluation benchmarks or suites (Guo et al., 2023; Liu et al., 2025), such as the GAOKAO benchmark (Zhang et al., 2023). The GAOKAO benchmark collects questions from past college entrance examination papers, gathering 936 questions by the year 2024. The questions are categorized into multiple-choice questions, fill-in-the-blank questions, and math word problems. Other benchmarks, like C-EVAL (Huang et al., 2023), M3KE (Liu et al., 2023), etc., broadly include various levels of math questions, such as elementary and advanced mathematics. These mathematical reasoning evaluation datasets suffer from three issues. First, the scale is relatively small; the number of problems is usually fewer than 1K. Second, the evaluation data are not meticulously categorized and divided, making the evaluation results less useful for understanding different mathematical reasoning abilities and levels. Third, unlike the MATH dataset, where all questions are math word problems, the above datasets mainly consist of single-choice and fill-in-the-blank questions, with a limited number of math word problems. We conduct a statistical analysis on the character length of problems and solutions in the above datasets. Results are shown in Table 1. The datasets that contain only multi-choice questions, such as C-EVAL and M3KE, have the shortest average question length. The average length of solutions in HighMATH is about twice that of GAOKAO, indicating that the problems in our dataset are more complex and require more reasoning steps.

Transformed Datasets. Currently, mathematical reasoning evaluations often use static benchmarks, which are prone to data contamination. To prevent or bypass this issue, efforts have been dedicated into (1) new datasets in different formats, (2) dynamic benchmarks and (3) synthetic datasets. MathVista collects multimodal math problems (Lu et al., 2024). Work (Zhu et al., 2023) attempt to develop dynamic test sets. However, unlike elementary mathematics, more challenging math problems are difficult to implement in dynamic evaluations. Other efforts (Mirzadeh et al., 2024; Gulati et al., 2024; Wei et al., 2023; Zhang and Xiong, 2025a,b; Wu et al., 2025) modify the original information in the questions of the GSM8K or other datasets, such as variable names or numeric values, using templates to generate variants of the original problems for evaluation. Among these methods, adding information unrelated to the questions can cause model performance to drop by as much as 65%. Our dataset shows that even with pure text, without data synthesis, LLMs still face challenges when performing mathematical reasoning.

3 Benchmark Curation

3.1 Motivation

Our work is motivated by the limitations of existing math benchmarks. Therefore, our design aims to create a dataset with appropriate difficulty, broad coverage of mathematical knowledge points, and fine-grained annotations. To this end, we decide to focus on math exams related to the Chinese college entrance, i.e., senior high school math exams.

Compared to elementary and middle school math, senior high school math covers knowledge about different mathematical concepts and presents a certain level of difficulty. We specifically collect questions that test Chinese high school senior students' mastery of comprehensive mathematical knowledge. These questions are often examining a variety of knowledge points within a single problem. Additionally, we choose math word problems as the test format. Although multiple-choice questions are easy to standardize and convenient for evaluation, the probability of selecting the correct answer is relatively high. Evaluation results hence may not accurately reflect the true reasoning capability of models. Therefore, during data collection, besides the original math word problems, we also convert valuable multiple-choice and fill-in-the-blank questions into math word problems for evaluation.

3.2 Data Collection

Subject Division. Based on the 2019 Chinese college entrance examination mathematics syllabus, we organize the collected data into eight subjects, which basically cover all the outlined knowledge points in the syllabus. The specific subjects and their corresponding mathematical contents are presented in Table 2.

Data Sources. HighMATH collects data from various sources, mainly from joint provincial examinations and mock exam papers from educational institutions published between January and May

Subject Category	Category Content	Size
Counting Principle	Permutations, combinations, binomial theorem, etc.	194
Logic Set Inequality Complex	Includes common operations, basic set operations, inductive proofs, linear programming, and proofs of inequalities, etc.	783
Function Derivative	Includes derivatives of one-variable functions and their applications, basic elementary functions and their properties, etc.	955
Solid Geometry	Relationships of points, lines, and planes in space, volumes of cylinders, spheres, inscribed spheres, circumscribed spheres, sectional problems, etc.	548
Trigonometric Functions and Triangle Solving	Includes trigonometric functions, trigonometric identity transformations, solving triangles, plane vectors, etc.	787
Plane Analytical Geometry	Includes conic sections, lines and circles, plane vectors, etc.	1081
Sequences	Arithmetic sequences, geometric sequences, etc.	609
Statistics Probability	Random variables and their distributions, random events and probabilities, distribution tables and hypergeometric distributions, data statistical analysis, etc.	336

Table 2: A comprehensive overview of mathematical subjects and their specific contents is provided. The number of problems each subject contains is listed in the Size column.

2024. Both of the selected papers are designed to assess the senior high school students' mastery of a variety of mathematical knowledge and reasoning capabilities. The problems in those papers are carefully selected or created by experienced teachers and exam setters to ensure quality. The diverse range of sources not only enhances the representativeness and challenge of the dataset, but also reflects the breadth of the college entrance examination. Moreover, to further reduce the risk of data contamination, we select papers saved in scanned PDF format rather than formats that can be directly crawled or parsed.

Data Processing. After collection, all files undergo automatic recognition followed by a review by a human annotator. Specifically, all the papers are first processed by OCR to recognize the problem and solution text, then math expressions are converted into LaTeX format using Mathpix. After that, the LaTeX expressions are compiled into a human-readable form. They are manually proofread for accuracy before being organized into the final standardized dataset. Annotators are also responsible for assigning difficulty levels ranging from 1 (easiest) to 7 (most difficult). All annotators are college students who have passed the college entrance examination and have rich experience and intuition with these problems. After the initial standardization of the data is completed, the leader of the annotators conducts multiple screenings and corrections to ensure annotation accuracy and consistency, thus ensuring the dataset's quality for efficient and precise use. The other detailed information on data processing, data review criteria, difficulty divisions, manual quality control, and HighMATH-HARD is listed in Appendix A. The annotated information and the final standardized format are illustrated in an example in Table 3.

3.3 Data Statistics

In order to demonstrate that HighMATH covers a rich variety of mathematical concepts and more knowledge points assessed by a single problem, we sample instances and conduct an analysis for comparison with MATH. The MATH dataset contains 7 categories and 5 difficulty levels, from which we randomly sample 70 problems. The HighMATH dataset includes 8 categories and 7 difficulty levels, resulting in 108 sampled problems. Annotators label the core knowledge points assessed by these sampled problems and the number of knowledge points that exist. After labeling, we analyze the results from three perspectives. First, we analyze the average number of knowledge points per problem. The MATH dataset averages 1.59 knowledge points per question, while the HighMATH dataset averages 2.4. Second, we evaluate the average number of knowledge points across subjects and difficulty levels. Results are shown in Figure 1. In both subject categories and difficulty levels, the average number of knowledge points in HighMATH exceeds those in the MATH dataset. Specifically, the average number of knowledge points is 2.07 times that of the MATH dataset for each subject, and

Information	Example in Chinese	English Translation
Question:	在 $\left(2x^3-\frac{1}{x}\right)^6$ 的展开式中, x^2 项的系数为?	In the expansion of $\left(2x^3-\frac{1}{x}\right)^6$, what is the coefficient of the x^2 term?
Solution:	【分析】由二项式展开式的通项公式写出其通项公式 $T_{k+1}=(-1)^k\times 2^{6-k}\times C_6^k\times x^{18-4k}$, 令 $18-4k=2$ 确定 k 的值, 然后计算 x^2 项的系数即可.【详解】展开式的通项公式 $T_{k+1}=\mathrm{C}_6^k\left(2x^3\right)^{6-k}\left(-\frac{1}{x}\right)^k=(-1)^k\times 2^{6-k}\times \mathrm{C}_6^k\times x^{18-4k}$,令 $18-4k=2$ 可得, $k=4$,则 x^2 项的系数为 $(-1)^4\times 2^{6-4}\times C_6^4=4\times 15=60$.	【Analysis】Using the general term formula of binomial expansion, write out the general term formula $T_{k+1}=(-1)^k\times 2^{6-k}\times C_6^k\times x^{18-4k}$, Set $18-4k=2$ to determine the value of k, then calculate the coefficient of the x^2 term. 【Detailed Solution】 The general term formula of the expansion is: $T_{k+1}=\mathrm{C}_6^k\left(2x^3\right)^{6-k}\left(-\frac{1}{x}\right)^k=(-1)^k\times 2^{6-k}\times \mathrm{C}_6^k\times x^{18-4k}$ Setting $18-4k=2$, we get $k=4$ Therefore, the coefficient of the x^2 term is: $(-1)^4\times 2^{6-4}\times C_6^4=4\times 15=60$.
Subject:	计数原理	Counting Principle
Level:	2	
Answer:	60	

Table 3: Illustration of annotated math word problems in HighMATH.

the average number of knowledge points is 1.69 times that of the MATH dataset for each difficulty level. Additionally, we conduct a manual evaluation to compare the breadth of knowledge points covered in HighMATH and MATH. We asked undergraduate students majoring in mathematics to evaluate the two datasets. The evaluation results show that, compared to MATH, HighMATH covers a more comprehensive range of knowledge, essentially encompassing all aspects of high school mathematics. Other evaluation opinions are shown in Appendix B. All manual annotations will be published along with the datasets.

4 Experiments

We conducted extensive experiments on HighMATH to evaluate 20 latest LLMs, including both open- and closed-source models. Among them, o1-like models were also evaluated. We divided all the problems from HighMATH into two categories for evaluation. The first category consists of problems that contain only one question, totaling 4,100 problems, and the second includes problems with multiple questions, totaling 1,193 problems (labeled "HighMATH-HARD" in Table 4). As shown in the results in Table 4.

4.1 Models

We evaluate four types of models: The first group of models includes open-source models designed for general purposes (Qwen-2.5-Instruct, Llama-3-Instruct); the second type comprises open-source LLMs that are fine-tuned for math reasoning (Qwen-2.5-Math, Qwen-2.5-Math-Instruct, deepseek-math-instruct, deepseek-math-rl, Mathstral-v0.1); for the third type, we evaluated some recently released o1-like open-source LLMs specifically designed for complex reasoning (Skywork-o1-Open-Llama-3.1, QwQ-32B-Preview, Qwen3-235b-a22b, DeepSeek-R1-Distill-Qwen, DeepSeek-R1-671B). Additionally, we evaluate closed-source models accessed through APIs (OpenAI o1-mini, GLM-Z1-AirX, hunyuan-t1-20250521, SparkX1, Doubao-1-5-thinking-pro-250415). We evaluated models of different sizes within the same open-source model series whenever possible, with model sizes ranging from 1.5B to 72B parameters. We follow the recommended settings outlined in the model documentation and configuration file for each tested model. For more detailed settings and prompt usage, please refer to Appendix C.

4.2 Evaluation Strategies

We employed four evaluation settings to provide a comprehensive analysis of model performance on HighMATH.

Zero-shot Evaluation: Zero-shot evaluation refers to the process where the model makes predictions on inputs and directly performs the evaluation task without having seen any specific samples before. During the evaluation, we only give the model a system prompt instructing it to write the final answer in a box, and directly provide the problem in the user prompt. This method is the most direct and widely used evaluation approach.

MODEL	HighMATH	HighMATH-HARD	Major Vote	Pass@8	Python Executor
	General Purpose Open-Source LLMs				
Qwen-2.5-Instruct-7B	21.71	5.87	43.34	57.00	-
Llama-3-Instruct-8B	6.49	1.34	8.90	20.73	-
	Mathematics Open-Source LLMs				
Qwen-2.5-Math-Instruct-1.5B	16.05	5.20	48.29	58.11	42.11
Qwen-2.5-Math-7B	9.15	0.42	28.83	42.05	-
Qwen-2.5-Math-Instruct-7B	43.71	18.44	51.76	60.44	46.41
Qwen-2.5-Math-Instruct-72B	51.56	24.48	-	-	50.62
deepseek-math-7b-instruct	15.73	3.10	19.22	35.93	-
deepseek-math-7b-rl	8.15	1.93	16.78	24.68	-
Mathstral-7B-v0.1	21.12	5.53	29.73	48.98	-
	o1-like Open-Source LLMs				
Skywork-o1-Open-Llama-3.1-8B	45.80	17.94	51.20	61.49	-
QwQ-32B-Preview	45.41	12.57	54.61	59.90	-
Qwen3-235b-a22b	70.07	-	-	-	-
DeepSeek-R1-Distill-Qwen-7B	29.68	5.53	43.44	45.07	-
DeepSeek-R1-Distill-Qwen-32B	31.22	6.87	45.53	47.24	-
DeepSeek-R1-671B	68.93	-	-	-	-
	Closed-Source LLMs				
o1-mini	52.53	-	-	-	-
GLM-Z1-AirX	65.73	-	-	-	-
hunyuan-t1-20250521	66.56	-	-	-	-
SparkX1	67.34	-	-	-	-
Doubao-1-5-thinking-pro-250415	71.02	-	-	-	-

Table 4: Main results. “HighMATH” and “HighMATH-HARD” are evaluated via zero-shot evaluation. Except for closed-source LLMs, other models have set the maximum generation length to 2048 for evaluation efficiency. We also list statistics on their completion of responses in Table 7 in the Appendix C for reference; if the “\boxed{}” marker is found in the response, it indicates that the model has completed the answer. “-” indicates that related experiments were not performed due to resource limitations.

Majority Vote Mechanism: The majority vote (Wang et al., 2022) mechanism involves selecting the most frequent answer from multiple inference responses (eight in our experiment) as the final decision, which is then compared to the ground truth. This approach helps assess the model's consistency and reliability in providing correct answers over multiple runs.

Pass@N: Pass@N is a metric used to evaluate the model's ability to find at least one correct answer within the top N most likely generated answers (eight in our experiment). It is commonly used to measure the model's recall capability and the diversity of generated responses. As an evaluation standard, Pass@N reveals the model's performance when generating multiple candidate answers, emphasizing its ability to cover a wide range of possibilities.

Solving with Python Executor: This evaluation method involves using the Python Executor to enhance the logic and computational accuracy of LLM's responses. During experiments, we evaluated Qwen models, which provide relevant interfaces using this method. Specifically, after generating a response, the model sends the response to a relevant agent, which then regenerates and provides the final response.

5 Main Result

The main result of our work is that we found models that originally achieved high accuracy on MATH all experienced significant drops in accuracy on our benchmark. For example, Qwen2.5-Math-1.5B/7B/72B-Instruct achieve 79.7%, 85.3%, and 87.8% respectively on the MATH benchmark using TIR (Yang et al., 2024), while Qwen2.5-Math-1.5B/7B/72B-Instruct only achieve 42.11%, 46.41%, and 50.62% respectively on our benchmark using TIR.

HighMATH can provide good discrimination among LLMs. Doubao-1-5-thinking-pro-250415 performs best in handling single-question problems, with an accuracy rate of 71.02%, which greatly surpasses other models. To ensure efficiency, we set a maximum generation length of 2048 tokens for all models except for closed-source LLMs. We discovered that R1's overly long reasoning process prevented it from completing the reasoning within the limit, thereby affecting its accuracy. After re-evaluating the model using its official recommended settings, the accuracy of DeepSeek-R1-Distill-Qwen-7B improved to 53.32%. It can be seen that Qwen-2.5-Instruct, Qwen-2.5-Math, and Qwen-2.5-Math-Instruct, which are based on the same 7B base model, show significant differences. Comparing Qwen-2.5-Instruct and Qwen-2.5-Math, it is evident that even though Qwen-2.5-Math has been trained with a substantial new mathematical corpus, its performance is greatly influenced by whether the model has undergone instruction-following tuning. Qwen-2.5-Math-Instruct, which combines the advantages of the two previous models, achieves the best results, indicating that both factors are important in model training. For Qwen-Math-Instruct models, as model size increases, performance on HighMATH does not improve a lot. Comparing the 7B model with the 72B model, there is only an 8% improvement in single-question problems' evaluation.

The accuracy of all models drastically decreases in the evaluation of multi-question problems. The best-performing model, Qwen-2.5-Math-Instruct, has an accuracy in multi-question problems that is less than half of its accuracy in single-question problems. Models that perform poorly on single-question problems tend to perform even worse on multi-question problems. Among the o1-like models, the accuracy of QWQ-32B-Preview on multi-question problems is surprisingly 5.87% lower than that of Qwen-2.5-Math-Instruct 7B. Besides Skywork-o1-Open-Llama-3.1-8B, other o1-like models also fail to meet the expected results, which may indicate that a more complex reasoning process does not necessarily improve accuracy.

The majority vote and pass@8 evaluation settings significantly improve the accuracy of each model in reasoning on single-question problems, although the degree of improvement varies. The most noticeable increases occur in models that originally have lower accuracies. For example, the pass@8 accuracy of the Qwen 1.5B model even surpasses that of many 7B models. These results may indicate that these smaller LLMs already possess strong mathematical reasoning abilities, but the responses they generate in a single pass are very unstable, leading to poor accuracy performance. Using a Python executor can effectively mitigate this issue. However, it is worth noting that there is an upper limit to the accuracy achievable with majority vote, pass@8, or using a Python executor on HighMATH. The highest accuracy is achieved by Skywork-o1-Open-Llama-3.1 under the pass@8 evaluation setting, reaching only 61.49%.

6 Analysis of Problem Subjects and Difficulty Levels

We conducted an analysis of performance across problem subjects and difficulty levels between o1-mini, QWQ-32B-Preview, Qwen-2.5-Math-Instruct-72B, Skywork-o1-Open-Llama-3.1-8B, and the Meta-Llama-3-8B-Instruct model. All comparisons are conducted under the evaluation results of single-question problems.

6.1 Problem Subjects

The left radargrame in Figure 2 shows that the compared models generally perform consistently across different categories of math problems. OpenAI o1-mini not only outperforms other models in overall accuracy on HighMATH but also performs best in five subjects of math problems, though it is slightly inferior in two categories. This indicates that OpenAI o1-mini is still demonstrates competitiveness in the current evaluation tasks. Clearly, compared to the Meta-Llama-3-8B-Instruct model, the Skywork-o1-Open-Llama-3.1-8B model shows significant improvements in accuracy across all categories. This finding again highlights that, with effective training and optimization strategies, even

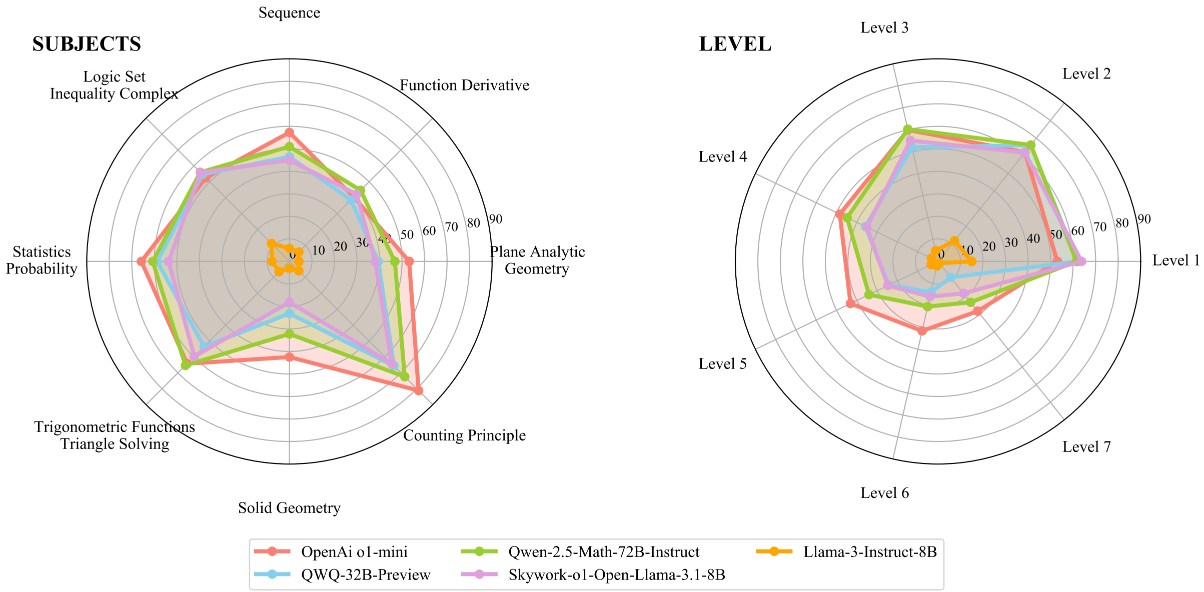

Figure 2: Analysis of question types and difficulty levels.

relatively smaller models can exhibit reasoning abilities on par with larger models.

We also observe that the models perform well in subjects such as Counting Principle, Sequences, and Statistics. We speculate that mathematical problems in these fields often exhibit strong structural characteristics, including clear rules and logic. The models can leverage a large number of mathematical formulas, derivation processes, and rules to perform calculations, making accurate reasoning more achievable. Furthermore, the richness of the training data and the standardized nature of the problems make these areas particularly advantageous for complex reasoning models to demonstrate their strengths. However, models perform poorly in subjects such as Solid Geometry and Plane Analytic Geometry. We believe that these subjects require the models to possess strong spatial imagination, complex geometric reasoning abilities, and a deep understanding of geometric transformations and shapes.

6.2 Difficulty Level

The accuracy of all models decreases as the difficulty level increases. OpenAI o1-mini surpasses other models in accuracy for problems at Level 4 and above, demonstrating its advantage in complex mathematical reasoning. Overall, performance differences below Level 3 are not significant. However, for more difficult problems, performance aligns with the scaling law principle. As model size increases (e.g., Qwen QWQ-32B-Preview, Qwen-2.5-Math-72B-Instruct), larger models progressively outperform smaller ones in high-difficulty problems, indicating a positive correlation between model size and performance on challenging tasks.

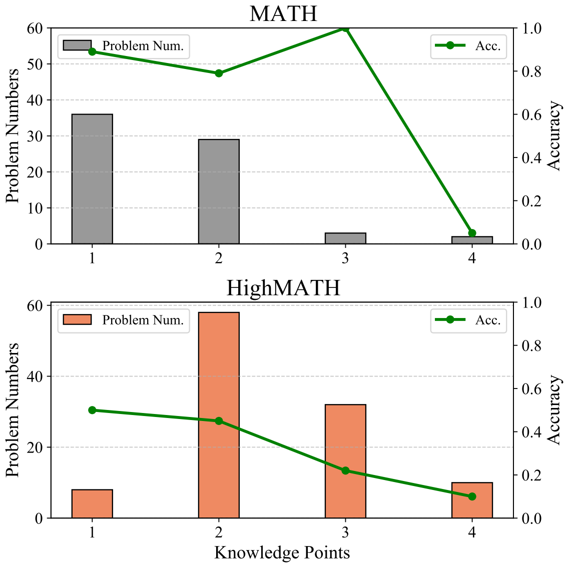

Figure 3: The accuracy of reasoning varies across problems with different numbers of knowledge points.

7 Challenges of LLMs in Associating Multiple Math Concepts

From the evaluation results presented above, it is evident that although the mathematical knowledge tested in both the MATH and HighMATH exams originates from the high school level, there is a sig-

Dataset	MATH	MATH500
Qwen-2.5-Math-Instruct-7B	76.6	75

Table 5: Evaluate Qwen-2.5-Math-Instruct-7b on MATH and MATH500.

Dataset	1000-Zh	1000-En
Qwen-2.5-Math-Instruct-7B	40.7	38.6

Table 6: The impact of language on mathematical evaluation.

nificant regression in performance on HighMATH. As demonstrated in Figure 1, HighMATH contains a higher average number of knowledge points compared to MATH. We conducted further analysis to examine whether LLMs face challenges in associating multiple math concepts. We used 70 samples from MATH and 108 samples from HighMATH, all of which have core knowledge points annotated as mentioned in Section 3.3. As seen in Figure 3, the range of the number of knowledge points annotated in both MATH and HighMATH varies between 1 and 4. In MATH, the majority of problems contain only one or two knowledge points, whereas in HighMATH, the majority of problems contain two or three knowledge points. In MATH, the accuracy of reasoning shows little variation across problems with one to three knowledge points. However, in HighMATH, the accuracy of reasoning varies more significantly across problems with different numbers of knowledge points, showing a clear downward trend as the number of knowledge points increases. This suggests that LLMs may indeed face difficulties in associating multiple math concepts, which also highlights the evaluative advantage of HighMATH.

8 Ablation Study

To demonstrate the effectiveness of our dataset, we conducted identical tests on the Qwen-2.5-Math model using both MATH and MATH-500. As shown in Table 5, both MATH and MATH-500 achieve a 75% accuracy under the zero-shot evaluation setting. Under the same settings, HighMATH achieves an accuracy of 43.71%, showing a significant difference of 31.29 percentage points compared to the other datasets. This difference validate the effectiveness of using HighMATH for evaluation.

To test whether language affects mathematical reasoning, we extracted 1,000 questions proportionally from the eight subjects of HighMATH and translated them into English using GPT-4-0613 for testing the Qwen-2.5-Math-Instruct-7B. Experimental results, shown in Table 6, indicate that the accuracy slightly decreases with the translated questions, but the change is not significant. Therefore, although HighMATH is based on a Chinese context, it remains relevant for testing models trained in other languages.

9 Conclusion

We propose a new mathematical reasoning dataset, HighMATH, for LLM evaluations. HighMATH features a comprehensive inclusion of different subjects of mathematical concepts, with multiple knowledge points integrated into each problem. Additionally, HighMATH includes fine-grained annotations, divided into 8 categories and 7 difficulty levels, totaling 5,293 problems. Based on HighMATH, we have conducted a comprehensive evaluation of the most recent popular LLMs. The evaluation results confirm the effectiveness of our dataset, with the best model achieving an accuracy of 62%.

Acknowledgments

The present research was supported by the National Key Research and Development Program of China (Grant No. 2023YFE0116400). We would like to thank the anonymous reviewers for their insightful comments.

Limitations

While our benchmark provides a comprehensive evaluation of LLMs' math reasoning abilities at the high school level, it has some limitations. First, it does not cover university-level mathematics or advanced Olympiad problems, focusing instead only on high school content. Additionally, considering the current progress in reasoning capabilities of models like o1 and DeepSeek-R1, we believe that it is equally crucial to evaluate each reasoning step generated by the model, not just focusing on the final answer. This may help to more accurately measure the model's performance in complex logical reasoning processes.

References

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024. Alphamath almost zero: process supervision without process. *arXiv preprint arXiv:2405.03553*.

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. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno de Moraes Dumont, and Sanmi Koyejo. 2024. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Leo Li, Ye Luo, and Tingyou Pan. 2024. Openai-o1 ab testing: Does the o1 model really do good reasoning in math problem solving? *arXiv preprint arXiv:2411.06198*.

Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, Xiaowen Su, Qun Liu, and Deyi Xiong. 2023. M3KE: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *arXiv preprint arXiv:2305.10263*.

Chuang Liu, Renren Jin, Zheng Yao, Tianyi Li, Liang Cheng, Mark Steedman, and Deyi Xiong. 2025. Empirical study on data attributes insufficiency of evaluation benchmarks for llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6024–6038.

Yan Liu, Renren Jin, Ling Shi, Zheng Yao, and Deyi Xiong. 2024. Finemath: A fine-grained mathematical evaluation benchmark for chinese large language models. *arXiv preprint arXiv:2403.07747*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *arXiv preprint arXiv:2408.06195*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. 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)*, pages 9426–9439.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. 2023. CMATH: Can your language model pass chinese elementary school math test? *arXiv preprint arXiv:2306.16636*.

Junru Wu, Tianhao Shen, Linxi Su, and Deyi Xiong. 2025. C^{2}RBench: A Chinese complex reasoning benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21031–21050, Vienna, Austria. Association for Computational Linguistics.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Shaowei Zhang and Deyi Xiong. 2025a. Backmath: Towards backward reasoning for solving math problems step by step. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 466–482.

Shaowei Zhang and Deyi Xiong. 2025b. Debate4MATH: Multi-agent debate for fine-grained reasoning in math. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16810–16824, Vienna, Austria. Association for Computational Linguistics.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2023. Dyval: Graph-informed dynamic evaluation of large language models. *arXiv preprint arXiv:2309.17167*.

10 Appendix

A Detailed Information on Data Processing

Nine annotators collected and annotated 7,223 problems over two months. For these problems, we first performed deduplication using the difflib.SequenceMatcher algorithm to remove duplicate problems (the threshold was set to 0.925 after testing). Next, we conducted some screening, which included checking if the annotations met the standards, removing irrelevant information from the problems, eliminating open-ended questions, and discarding problems with incomplete information. These steps eliminated 1,320 problems, leaving 5,903. Finally, through real evaluation (using the MetaMath evaluation script), we removed problems that included Chinese descriptions in the answers and those with improperly marked golden answers, based on whether the evaluation script could correctly extract the golden answer. This step removed 610 problems, leaving a final total of 5,293 problems.

A.1 Data Review Criteria

According to our data annotation manual, annotators first need to compare the information from automatic recognition with the content in the PDF to ensure the completeness of the problem. Annotators also need to confirm whether the problem is usable. For example, if a solid geometry problem includes both text and image descriptions, annotators must check that the information in the image is already included in the text description and that the question can be answered without the image; otherwise, the problem will be discarded. For problems that need modification, annotators must simultaneously modify the question and its solution to ensure consistency and smoothness. Most importantly, annotators must confirm the correctness of the LaTeX code for mathematical formulas.

A.2 Difficulty Division

We follow three criteria for assessing difficulty:

- We use the difficulty level of the college entrance examination in China as a reference standard. We consider the hardest problems in the college entrance exam as difficulty level 7, and the basic problems, such as the first problem in multiple-choice or fill-in-the-blank sections, as difficulty level 1. Additionally, college entrance exam simulation papers usually have a fixed structure, where the last problem is typically the hardest.

- We assume that the longer the solution steps for a problem, the relatively harder it is.

- The determination of difficulty is based on the comparison of problems within the same category, rather than entirely on the subjective judgment of the annotators.

A.3 Selection Criteria for HighMath-HARD

HighMATH-HARD selects some of the last problems listed in the exam papers we have collected. Those exam papers usually follow a fixed structure, with the last few problems typically being the hardest and containing multiple sub-questions. The problems in HighMATH-HARD each contain at least two sub-questions. Compared to problems that contain only one question, those in HighMATH-HARD better evaluate the model's complex reasoning abilities because these problems are interconnected. The model can only solve the next question based on solving the previous one, and it is considered correct only if it completely answers the last sub-question correctly. This poses very high demands on the model's capabilities.

B Human Evaluation Opinions on HighMATH and MATH

- Problems in MATH are simpler and more direct, lacking integration of knowledge points, which means they do not combine two or more knowledge points for examination.

- Problems in HighMATH are more obscure, requiring people to engage in logical thinking to understand the meaning of the questions.

- The HighMATH dataset covers a more comprehensive range of knowledge, essentially encompassing all aspects of high school mathematics.

C Experiment Settings

During the evaluation, we follow the recommended settings outlined in the model documentation and configuration file for each tested model. We have set a uniform maximum generation length (max_tokens=2048) for open-source LLMs. Statistics on their completion of responses can be found

Models	\boxed{} Num	Acc
Qwen-2.5-Instruct-7B	4140	21.71
Llama-3-Instruct-8B	3982	6.49
Qwen-2.5-Math-Instruct-1.5B	3902	16.05
Qwen-2.5-Math-7B	1637	9.15
Qwen-2.5-Math-Instruct-7B	4105	43.71
Qwen-2.5-Math-Instruct-72B	4305	51.56
deepseek-math-7b-instruct	4116	15.73
deepseek-math-7b-rl	1944	8.15
Mathstral-7B-v0.1	4398	21.12
QwQ-32B-Preview	2929	45.41
Skywork-o1-Open-Llama-3.1-8B	4096	45.80
DeepSeek-R1-Distill-Qwen-7B	1768	29.68
DeepSeek-R1-Distill-Qwen-32B	1856	31.22
DeepSeek-R1-Distill-Qwen-7B w/o max_tokens	3673	53.32

Table 7: Statistics on models' response completion.

Difficulty	Count
Level 1	49
Level 2	112
Level 3	167
Level 4	250
Level 5	199
Level 6	112
Level 7	46

Table 8: The incorrectly answered problems cover various difficulty levels.

Subject	Count
Counting Principle	19
Function Derivative	189
Logic Set Inequality Complex	135
Plane Analytic Geometry	187
Sequence	127
Solid Geometry	165
Statistics Probability	35
Trigonometric Functions and Triangle Solving	78

Table 9: The incorrectly answered problems cover various categories.

in table 7 (if \boxed{} is found in the response, it indicates that the model has completed the answer). We use the evaluation code open-sourced by MetaMath (Yu et al., 2023) to ensure uniform evaluation rules.

The two formats of prompts that we use are listed below, and we choose which to use based on the recommendations in the model documentation.

messages = ["role": "system", "content": "Please reason step by step, and put your final answer within \boxed{}.", "role": "user", "content": problem]

messages = ["role": "user", "content": problem + "\n" + "Please reason step by step, and put your final answer within \boxed{}."]

D Extended Error Analysis

We have started to compare the results of all the models we have evaluated (we chose the evaluation results of pass@8 with the highest accuracy for this analysis). The results show that 935 problems were not answered correctly by any model. These problems cover various difficulty levels and categories, as shown in Table 8 and Table 9.