U-MATH: A University-Level Benchmark for Evaluating Mathematical Skills in Large Language Models

Konstantin Chernyshev[*], **Vitaliy Polshkov, Ekaterina Artemova, Sergei Tilga**
Toloka AI
`{kchernyshev, cogwheelhead, katya-art, tilgasergey}@toloka.ai`

Alex Myasnikov, Vlad Stepanov
Gradarius
`{alex, vstepanov}@gradarius.com`

Alexei Miasnikov
Gradarius, Stevens Institute of Technology
`amiasnik@stevens.edu`

Abstract

Current evaluations of mathematical skills in Large Language Models are constrained by benchmarks lacking scope, particularly for multi-modal problems — frequently relying on school-level (Cobbe et al., 2021; Lu et al., 2023; Zhang et al., 2024), niche Olympiad-style (Fang et al., 2024; Mao et al., 2024), simple quiz format (Yue et al., 2023; Qiao et al., 2024) or relatively small (Lewkowycz et al., 2022) datasets.

To address this, we introduce **U-MATH**, a novel benchmark comprising **1,100** unpublished open-ended university-level problems sourced from current US curricula, with **20%** incorporating visual elements. Given the free-form nature of U-MATH problems, we employ LLM judges for solution evaluation and release **μ-MATH**, a meta-evaluation benchmark composed of **1,084** U-MATH-derived tasks enabling precise assessment of these judges.

Benchmarking leading LLMs reveals marked limitations in multi-modal reasoning, with maximum accuracy reaching 93.1% on textual tasks but only 58.5% on visual ones. Furthermore, solution judgment proves challenging, requiring the most advanced models to achieve meaningfully high performance, even still peaking at an imperfect F1-score of 90.1%.

We open-source U-MATH, μ-MATH, and all our evaluation code.[1]

1 Introduction

Assessing the mathematical proficiency of Large Language Models (LLMs) is crucial for evaluating their fundamental reasoning capabilities (Ahn et al., 2024). The most widely used benchmarks, GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), primarily cover school-level problems, overlooking advanced topics and facing rapid saturation (Achiam et al., 2023). Although some MATH problems and other recent works introduce harder concepts, they are limited in size and scope, relying on competition-style problems and neglecting the practical middle-ground of university-level coursework.

There is also growing demand for visual reasoning assessment in multi-modal LLMs (Ahn et al., 2024). Datasets such as the recent MATH-V (Wang et al., 2024a) provide numerous visual problems but face similar topic limitations or rely on the multiple-choice format, making the tasks significantly easier (Li et al., 2024b; Pezeshkpour and Hruschka, 2023).

In turn, reliably evaluating complex free-form responses is challenging (Hendrycks et al., 2021), which results in LLM judges becoming the de facto standard despite known biases and inconsistencies (Zheng et al., 2023). These biases are often overlooked and unquantified, preventing potential correction. Quantifying auto-evaluation errors requires datasets designed specifically to assess the evaluators themselves, also called meta-evaluations. While mathematical meta-evaluation datasets do exist, they are mostly based on GSM8K and MATH, inheriting their scope limitations.

To address these gaps, we introduce the **U-MATH** (*U*niversity *Math*) and **μ-MATH** (*M*eta *U-MATH*) benchmarks. Our main contributions are:

1. **U-MATH** (Section 3): We open-source 1,100 university-level problems, balanced across six core university subjects. The problems are collected from actual coursework and supplied with correct answers, with approximately 20% incorporating visual elements.
2. **μ-MATH** (Section 3.3): We introduce a set of 1084 meta-evaluation tasks designed to assess the quality of LLM judges by selecting

---

[*] Corresponding author: `kchernyshev@toloka.ai`

[1] `https://github.com/toloka/u-math`

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 974–1001
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Example: Differential Calculus.

U-MATH Problem:
The function $s(t) = 2 \cdot t^3 - 3 \cdot t^2 - 12 \cdot t + 8$ represents the position of a particle traveling along a horizontal line.
1. Find the velocity and acceleration functions.
2. Determine the time intervals when the object is slowing down or speeding up.

Reference Solution (shortened):
The velocity is $v(t) = s'(t) = \boxed{6 \cdot t^2 - 6 \cdot t - 12}$, zeros of the $v(t)$ are $t = -1, 2$.
The acceleration is $a(t) = v'(t) = \boxed{12 \cdot t - 6}$, zero of the $a(t)$ is $t = \frac{1}{2}$.
It speeds up when $v(t)$ and $a(t)$ have the same sign, and slows down when opposite.

Interval	$v(t)$	$a(t)$	Behavior
$(-\infty, -1)$	> 0	< 0	Slowing down
$(-1, \frac{1}{2})$	< 0	< 0	Speeding up
$(\frac{1}{2}, 2)$	< 0	> 0	Slowing down
$(2, \infty)$	> 0	> 0	Speeding up

Accounting for non-negative time, speed up on $\boxed{(0, 1/2) \text{ and } (2, \infty)}$, slow down on $\boxed{(1/2, 2)}$.

Figure 1: A U-MATH sample. A common students' error reported by the author is overlooking time non-negativity.

approximately 25% of the U-MATH problems, supplying each with four solutions produced by four different top-performing language models, and providing ground truth labels on generated solutions' correctness.

3. **Comparative analysis** (Section 4): We compare various open-source and proprietary LLMs on U-MATH and μ-MATH, revealing significant deficiencies in solving university-level multi-modal problems. We also find proprietary models to outperform open-source ones on these tasks, while near-parity is observed with the text modality. Judgment also proves challenging for LLMs, with only the best-performing and most recent models attaining adequately high scores. In addition, we demonstrate that most current systems exhibit biased and unstable judgment performance. Finally, we establish that judgment as a skill is distinct from problem-solving and identify characteristic behavioral tendencies in LLM judges.

We release the U-MATH and μ-MATH benchmarks under a permissive license to facilitate further research and ensure reproducibility.

2 Background

Evaluating mathematical capabilities of LLMs is an essential direction of AI research (Ahn et al., 2024). Apart from mathematical proficiency being important in and of itself, studies show that fine-tuning with math and code-related data enhances models' fundamental 'cognitive skills' (Prakash et al., 2024) and reasoning capabilities (Chen et al., 2024), further necessitating the creation of mathematical evaluation datasets. Despite significant progress, many existing datasets are limited in scope, complexity of the problems, or size, as evidenced by the summary in Table 1.

Textual Mathematical Benchmarks. Datasets like MathQA (Amini et al., 2019) and the mathematics subset of MMLU (Hendrycks et al., 2020) represent early efforts to assess math capabilities of LLMs, relying primarily on rather simple multiple-choice problems. Today, even smaller models have achieved high scores with these tasks (Li et al., 2024a), rendering the benchmarks obsolete.

Subsequently, more comprehensive datasets emerged, including GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), and MGSM (Shi et al., 2022) (a multilingual version of 250 GSM8K samples). These, however, mostly include elementary- to high-shool level problems, which may not fully gauge the depth of mathematical reasoning, and quickly approach saturation as well.

Recent works aim to introduce more advanced concepts, prominent examples including MathOdyssey (Fang et al., 2024) and CHAMP (Mao et al., 2024), composed primarily of problems from high-school competitions, ProofNet (Azerbayev et al., 2023) and MiniF2F (Zheng et al., 2021), focused on formal proof composition and autoformalization, and OCWCourses (Lewkowycz et al., 2022), based on MIT curricula contents. However, these datasets are constrained by their

Dataset	Levels	%Uni. Level	#Test	%Visual	%Free-form	#Free-form Text-only Uni. Level Test	#Free-form Visual Uni. Level Test
MMLU$_{Math}$ (Hendrycks et al., 2020)	E H C	0	1.3k	0	0	0	0
GSM8k (Cobbe et al., 2021)	E	0	1k	0	0	0	0
MATH (Hendrycks et al., 2021)	H O	0	5k	0	100	0	0
MiniF2F (Zheng et al., 2021)	E H O	0	244	0	100	0	0
OCWCourses (Lewkowycz et al., 2022)	U	100	272	0	100	272	0
ProofNet (Azerbayev et al., 2023)	C U	≈50	371	0	100	≈180	0
CHAMP (Mao et al., 2024)	H O	0	270	0	100	0	0
MathOdyssey (Fang et al., 2024)	H U O	≈25	387	0	100	≈50	0
MMMU$_{Math}$ (Yue et al., 2023)	C	0	505	100	0	0	0
MathVista (Lu et al., 2023)	E H C	0	5k	100	46	0	0
MATH-V (Wang et al., 2024a)	E H O	0	3k	100	50	0	0
We-Math (Qiao et al., 2024)	E H U	≈20	1.7k	100	0	0	0
MathVerse (Zhang et al., 2024)	H	0	4.7k	83.3	45	0	0
U-MATH (this work)	U	100	1.1k	20	100	**900**	**200**

Table 1: Existing auto-evaluated math benchmarks along with their sizes, visual sample percentages, and open-ended problem percentages. Level markers: E Elementary to Middle School, H High School, C College, U University, O Olympiads.

smaller sizes (under 400 problems each), and most focus on Olympiad-style problems, missing the more practical topics of university coursework. Apart from that, all of them rely on publicly available materials, allowing for data leakage.

Our dataset offers **over three times more open-ended university-level problems** compared to these existing alternatives, with all of its problems previously unpublished.

Visual Mathematical Benchmarks. With the rise of multi-modal LLMs, demand for visual mathematical benchmarks is growing (Zhang et al., 2024; Qiao et al., 2024). Early efforts focused primarily on simpler geometry problems, as seen with datasets such as GeoQA (Chen et al., 2022b), UniGeo (Chen et al., 2022a), and Geometry3K (Lu et al., 2021), which offer a very narrow coverage of visual reasoning.

Later developments attempted to broaden the scope. MMMU (Yue et al., 2023) provides 505 college-level visual questions, but its complexity is limited by the use of multiple-choice format. MathVista (Lu et al., 2023) combines 28 existing and 3 new datasets, totaling 5k samples (1k test), although Qiao et al. (2024) noted issues with data quality.

The latest benchmarks face similar limitations. We-Math (Qiao et al., 2024) includes 1.7k visual samples but again only uses the multiple-choice format. MathVerse (Zhang et al., 2024) and MATH-V (Wang et al., 2024a) both incorporate over 1.5k free-form solutions, but lack topic coverage due to their focus on simpler problems or high-school competition challenges.

Our U-MATH$_{\text{Visual}}$ subset embraces the **free-form response format for visual problems** while adhering to the topics of **university coursework**.

Mathematical solution verification. The open-ended nature of answers and ambiguity in mathematical expressions make evaluating math solutions particularly challenging. As a result, many benchmarks use multiple-choice questions for ease of grading, though this can simplify the tasks and offer hints that models can exploit (Li et al., 2024b; Pezeshkpour and Hruschka, 2023).

Free-form evaluation by LLM judges, while widespread (Zheng et al., 2023), is prone to errors that are often overlooked and unaccounted for, compromising reliability (Zheng et al., 2023). Therefore, tools allowing for assessment of automatic evaluators — meta-evaluations — are crucial. Recent studies also indicate that evaluating math solutions is challenging for LLMs (Zeng et al., 2023; Xia et al., 2024) and that judgment performance correlates with problem-solving performance without fully aligning with it (Stephan et al., 2024), further reinforcing the relevance of meta-evaluations.

There are existing datasets suited for mathematical meta-evaluations: PRM800K (Lightman et al., 2023) contains 800K annotated steps from 75K solutions to 12K MATH dataset problems, FELM (Zhao et al., 2024) provides GPT-3.5 annotations for solutions to 208 GSM8K and 194 MATH problems, MR-GSM8K (Zeng et al., 2023) and MR-MATH (Xia et al., 2024) introduce meta-evaluation tasks based on the problems from GSM8K and MATH. These are all essentially based on GSM8K and MATH datasets, neglecting meta-evaluation for more advanced mathematical areas.

Our μ-MATH benchmark is based on U-MATH problems, enabling **university-level meta-evaluations**.

3 U-MATH

We present **U-MATH** — a benchmark of 1,100 problems designed to evaluate LLMs' proficiency in university-level mathematics. Following prior work (Hendrycks et al., 2020, 2021; Cobbe et al., 2021; Fang et al., 2024; Yue et al., 2023), we use **Accuracy** as our main performance metric, employing an LLM judge (Zheng et al., 2023) to test evaluated responses against the golden labels. A problem is only considered solved if each of the questions included with the problem statement is answered correctly and fully (e.g. if one of the questions asks to find the saddle points of a function, all of them have to be found).

3.1 Dataset Curation

We collaborate with Gradarius, a platform providing math-specialized learning content and software for top US universities, sourcing tens of thousands of problems from ongoing courses across various institutions. Both problems and solutions are crafted by subject matter experts, representing real-world academic standards, and have not been externally published prior to our work. To build our benchmark, we select the most challenging problems available. In particular, we seek to filter out any calculation-intensive problems and focus on evaluating reasoning rather than arithmetical aptitude, as LLMs are not designed to perform arithmetic and are inherently prone to errors (Hendrycks et al., 2021; Lewkowycz et al., 2022).

First, we filter out problems with short solutions ($<$ 100 characters), problems in multiple-choice format, and problems marked as implying calculator use. Additionally, for visual problems, we choose to keep only those containing a single image, for evaluation simplicity.

Next, we employ several small language models — Llama-3.1 8B (Dubey et al., 2024), Qwen2 7B (Yang et al., 2024a), Mistral 7B (Jiang et al., 2023), Mathstral 7B, NuminaMath 7B (Beeching et al., 2024) — to solve the problems and select 150 most challenging ones per subject, based on the average solution rate. By using a diverse set of model families, we avoid allowing any individual one to be overly influential in problem selection.

Lastly, we manually curate the selected problems using our in-house mathematical experts and the Gradarius content team to ensure the absence of erroneous problem statements or golden labels.

Following the data curation, we enlist a team of academic experts from the Stevens Institute of Technology, who actively teach various Calculus courses. These experts thoroughly review the problems to verify whether they are suitable for assessing the subject knowledge expected of university students. Overall, only 4.3% of the problems are categorized as high-school rather than university-level.

3.2 Dataset Statistics

The U-MATH benchmark comprises **1,100** mathematical problems spanning **6 subjects**, with about **20%** of the problems including visual elements (graphs, tables, geometric figures). Table 2 summarizes the problems' distribution across the subjects, together with the average number of questions posed and answers expected per problem (e.g. the task could be to find the local minima, maxima, and saddle points of a function, while the correct answer might contain no extrema and two saddle points).

Math Subject	**#Textual**	**#Visual**	**Avg. Questions**	**Avg. Answers**
Algebra	150	30	1.93	1.28
Differential Calculus	150	70	2.37	1.15
Integral Calculus	150	58	1.09	1.01
Multivariable Calculus	150	28	1.74	1.09
Precalculus	150	10	1.51	1.23
Sequences and Series	150	4	1.36	1.00
All	900	200	1.66	1.12

Table 2: Statistics across U-MATH subjects: counts of text-only and visual problems, average questions per problem, and average answers per question

3.3 Meta-Evaluation Framework (μ-MATH)

Evaluating mathematical problems is not straightforward, with even simple expressions such as $x \cdot 0.5$ having alternative valid forms such as $\frac{x}{2}$, $x \div 2$, $x/2$, or unsimplified variants like $9x/18$. In practice, evaluating free-form solutions requires testing expression equivalence in much less trivial cases, especially with more advanced problems (see Appendix A.3 for an example). To systematically study the ability of LLMs to evaluate free-form mathematical solutions on advanced university-level problems, we introduce the **μ-MATH** benchmark. It consists of a curated subset of U-MATH samples, supplied with LLM-generated solutions, both correct and not. Four solutions are generated for each of the problems — using Qwen2.5 72B, Llama-3.1 8B, GPT-4o and Gemini 1.5 Pro models. We focus on text-only problems due to the limited size of the U-MATH$_{\text{Visual}}$ subset.

Solution correctness is determined using a combination of manual labeling and automatic verification via Gradarius-API, which allows to test formal equivalence of mathematical expressions. Whenever the API classifies an LLM-produced answer as coinciding with the golden label, we can be confident in that answer's correctness. However, a negative API response does not imply incorrectness, since extraction of the answer from the full solution and its subsequent conversion into an API-compatible expression format are imperfect. Hence, solutions with negative API responses, which occur roughly 40% of the time, are labeled by in-house math experts, same as described in Section 3.1.

Our internal experts also review all the problems, including the ones with all the solutions auto-labeled, to assess their evaluation difficulty. In the end, we select **271 U-MATH problems** (around **25%**) based on these difficulty estimates, resulting in a total of **1,084 samples**. The final set does not aim to reflect the overall U-MATH distribution, but rather provide a robust and challenging test for LLM judges.

A tested model is provided with a problem statement, a reference answer, and a solution to evaluate and is expected to produce a correctness judgment to be compared against the golden verdict. We treat this as a binary classification task, using the **macro-averaged F1-score** as our primary metric. To offer a finer-grained evaluation, we also report Positive Predictive Value (PPV or Precision) and True Positive Rate (TPR or Recall) for the positive class, as well as Negative Predictive Value (NPV) and True Negative Rate (TNR) for the negative class. We report scores calculated both overall (all samples) and per originating model, separately for each of the four author models.

4 Experiments and Results

4.1 Experimental Setup

We select some of the recent top-performing LLMs to evaluate (Table 3). All the non-reasoning models are restricted to a single generation of 4,096 tokens with temperature set to 0.

For reasoners, the token limit is 32,768. Note that o-series models do not allow for inference temperature control, always having a default nonzero temperature. Our internal tests on a subset of the models, including DeepSeek-R1 and QwQ-32B-Preview for the reasoner subset, show negligible differences in accuracy under greedy decoding and four-rollout Pass@1 with a temperature of 0.6 (average accuracy over four independent launches), nor do we observe any significant variation across the rollouts. We thus adhere to a single-generation scheme for reasoners as well, employing greedy decoding for all the models except the o-series.

Model	Source	Size(s)	Visual	Open-weights	Reasoner
Ministral 2410	Mistral.ai (2024a)	8B	✗	✓	✗
Mistral Small 2501	Mistral.ai (2024c)	24B	✗	✓	✗
Mistral Large 2411	Mistral.ai (2024b)	123B	✗	✓	✗
DeepSeek-V3	DeepSeek-AI et al. (2024)	MoE 37/685B	✗	✓	✗
Qwen2.5-Math	Yang et al. (2024b)	7B, 72B	✗	✓	✗
Qwen2.5	Team (2024)	7B, 32B, 72B	✗	✓	✗
Athene-V2 Chat	Nexusflow (2024)	72B	✗	✓	✗
Llama-3.1	Dubey et al. (2024)	8B, 70B	✗	✓	✗
Llama-3.1 Nemotron	Wang et al. (2024b)	70B	✗	✓	✗
Llama-3.3	Wang et al. (2024b)	70B	✗	✓	✗
Pixtral 12B 2409	Mistral AI (2024)	12B	✓	✓	✗
Pixtral Large 2411	Mistral AI (2024)	124B	✓	✓	✗
Qwen2-VL	Yang et al. (2024a)	7B, 72B	✓	✓	✗
Llama-3.2	Meta AI (2024)	11B, 90B	✓	✓	✗
Claude 3.5 Sonnet (new)	Anthropic (2024)	unknown	✓	✗	✗
GPT-4o-mini-2024-07-18	OpenAI (2024a)	unknown	✓	✗	✗
GPT-4o-2024-08-06	OpenAI (2024a)	unknown	✓	✗	✗
Gemini 1.5 Flash 002	Team et al. (2024)	unknown	✓	✗	✗
Gemini 1.5 Pro 002	Team et al. (2024)	unknown	✓	✗	✗
DeepSeek-R1	DeepSeek-AI et al. (2025)	MoE 37/685B	✗	✓	✓
QwQ-Preview	QwenLM (2024b)	32B	✗	✓	✓
QVQ-Preview	QwenLM (2024a)	72B	✓	✓	✓
o1-mini-2024-09-12	OpenAI (2024c)	unknown	✗	✗	✓
o3-mini-2025-01-31	OpenAI (2024d)	unknown	✗	✗	✓
o1-2024-12-17	OpenAI (2024b)	unknown	✓	✗	✓
Gemini 2.0 Flash Thinking (exp-01-21)	Google (2024)	unknown	✓	✗	✓

Table 3: The LLMs used in our work.

We use chain-of-thought prompting (Wei et al., 2022) with the prompt provided in Appendix C.1 and o3-mini as a judge, due to the model being simultaneously one of the most performant and balanced judges according to our meta-evaluations (see Section 4.3), as well as cost-effective and widely available, allowing for easier reproduction.

4.2 U-MATH Results

Table 4 summarizes the results of our experiments. We observe several key trends.

Reasoners offer breakthrough performance: Reasoning models attain the top U-MATH, U-MATH$_\mathrm{T}$ and U-MATH$_\mathrm{V}$ scores of 86.8%, 93.1% and 58.5% respectively, compared to 67.2%, 71.7% and 47.0% for the standard-inference models.

Open models are catching up on text-only problems, with DeepSeek in the lead: DeepSeek-V3 achieves a U-MATH$_\mathrm{T}$ score of 69.3%, closely trailing the leading Gemini 1.5 Pro model with 71.7%. DeepSeek-R1 (91.3%) is only marginally behind o1, the best-performing reasoner (93.1%).

Open models lag behind in visual problems, where Gemini dominates: The open-proprietary gap becomes much more pronounced when considering U-MATH$_\mathrm{V}$. In each 'capability group' (smaller, larger, and reasoning models) the best open-weight result comes from the Qwen family (Qwen2-VL 7B: 27.1%, Qwen2-VL 72B: 43.9%,

Model	U-MATH	U-MATH		Algebra		Diff. C.		Integral C.		Multivar C.		Precalculus		Seq.& Series	
		T 900	**V** 200	**T** 150	**V** 30	**T** 150	**V** 70	**T** 150	**V** 58	**T** 150	**V*** 28	**T** 150	**V*** 10	**T** 150	**V*** 4
Text-only models															
Ministral 8B	23.1	26.9	6.0	60.0	6.7	13.3	8.6	10.0	5.2	12.7	3.6	47.3	0.0	18.0	0.0
Llama-3.1 8B	29.5	33.7	11.0	60.0	3.3	17.3	10.0	22.7	19.0	23.3	3.6	50.7	20.0	28.0	0.0
Qwen2.5 7B	43.3	50.4	11.0	86.0	20.0	30.7	4.3	32.0	19.0	36.7	3.6	78.7	10.0	38.7	0.0
Qwen2.5-Math 7B	45.5	53.0	11.5	84.7	6.7	32.0	8.6	24.0	17.2	44.0	10.7	81.3	0.0	52.0	50.0
Mistral Small (24B)	34.8	39.9	12.0	80.7	13.3	13.3	10.0	13.3	15.5	25.3	14.3	70.7	0.0	36.0	0.0
Qwen2.5 32B	52.4	60.4	16.0	92.0	13.3	42.7	11.4	34.7	25.9	50.0	17.9	85.3	0.0	58.0	0.0
Llama-3.1 70B	35.2	40.4	11.5	79.3	3.3	17.3	17.1	16.0	10.3	26.7	7.1	68.0	0.0	35.3	50.0
Llama-3.1 Nemotron 70B	42.5	47.7	19.5	84.0	**23.3**	29.3	21.4	21.3	19.0	40.7	14.3	67.3	20.0	43.3	0.0
Llama-3.3 70B	44.7	51.7	13.5	83.3	6.7	35.3	11.4	27.3	20.7	48.7	10.7	68.7	10.0	46.7	25.0
Qwen2.5 72B	51.2	58.9	16.5	90.7	16.7	36.7	15.7	35.3	17.2	52.0	14.3	84.0	10.0	54.7	50.0
Athene-V2 Chat (72B)	54.9	62.9	19.0	87.3	10.0	43.3	22.9	36.7	17.2	62.0	21.4	**90.7**	0.0	57.3	**75.0**
Qwen2.5-Math 72B	59.5	68.7	18.0	94.7	6.7	46.0	12.9	**44.0**	25.9	**69.3**	21.4	89.3	10.0	68.7	**75.0**
Mistral Large (123B)	47.6	55.6	12.0	85.3	13.3	32.0	8.6	36.7	15.5	45.3	14.3	78.0	0.0	56.0	25.0
DeepSeek-V3 (MoE 37/685B)	**62.6**	**69.3**	**32.5**	**96.0**	10.0	**49.3**	**30.0**	38.7	**39.7**	**69.3**	**42.9**	90.0	**40.0**	**72.7**	50.0
Multimodal models															
Pixtral 12B	17.5	17.9	16.0	40.0	23.3	10.7	30.0	4.7	3.4	6.7	7.1	32.0	0.0	13.3	0.0
Llama-3.2 11B	20.4	22.9	9.0	52.0	3.3	7.3	20.0	1.3	3.4	13.3	0.0	44.0	10.0	19.3	0.0
Qwen2-VL 7B	26.3	27.1	22.5	58.7	10.0	18.7	37.1	11.3	17.2	14.0	17.9	42.7	10.0	17.3	0.0
Llama-3.2 90B	37.2	41.8	16.5	82.0	23.3	21.3	27.1	11.3	5.2	30.0	10.7	70.0	0.0	36.0	25.0
Qwen2-VL 72B	41.8	43.9	32.5	80.0	26.7	29.3	44.3	22.0	27.6	32.0	28.6	66.0	10.0	34.0	25.0
Pixtral Large (124B)	47.8	51.4	31.5	82.7	33.3	30.0	32.9	24.7	**32.8**	46.7	28.6	73.3	30.0	51.3	0.0
Claude Sonnet 3.5	38.7	40.7	30.0	75.3	30.0	20.7	41.4	12.0	15.5	33.3	39.3	64.0	20.0	38.7	0.0
GPT-4o-mini	43.4	47.2	26.0	87.3	13.3	26.0	32.9	16.7	17.2	37.3	39.3	76.0	20.0	40.0	50.0
GPT-4o	50.2	53.9	33.5	90.0	33.3	30.0	37.1	27.3	27.6	49.3	42.9	80.0	30.0	46.7	0.0
Gemini 1.5 Flash	57.8	61.2	42.5	90.7	46.7	47.3	47.1	30.7	31.0	55.3	53.6	82.7	30.0	60.7	50.0
Gemini 1.5 Pro	**67.2**	**71.7**	**47.0**	**92.0**	**60.0**	**62.0**	**50.0**	**47.3**	27.6	**65.3**	**60.7**	**90.0**	**50.0**	**73.3**	**75.0**
Reasoning models															
QVQ-72B-Preview	65.0	69.7	44.0	94.0	33.3	54.0	41.4	41.3	55.2	65.3	50.0	95.3	30.0	68.0	0.0
QwQ-32B-Preview	73.1	82.7	30.0	95.3	3.3	70.0	24.3	67.3	50.0	80.7	32.1	97.3	20.0	85.3	50.0
DeepSeek-R1 (MoE 37/685B)	80.7	91.3	33.0	96.7	16.7	85.3	22.9	87.3	50.0	86.7	42.9	98.7	10.0	93.3	**75.0**
o1-mini	76.3	82.9	46.5	97.3	40.0	75.3	52.9	72.0	46.6	78.7	42.9	96.7	30.0	77.3	50.0
Gemini 2.0 Flash Thinking	83.6	89.2	**58.5**	95.3	**60.0**	80.7	48.6	88.7	**65.5**	85.3	**75.0**	95.3	**50.0**	90.0	25.0
o3-mini	82.2	92.8	34.5	**99.3**	10.0	**88.0**	17.1	**90.7**	60.3	85.3	50.0	**99.3**	20.0	**94.0**	**75.0**
o1	**86.8**	**93.1**	**58.5**	97.3	50.0	86.0	**57.1**	**90.7**	63.8	**92.0**	60.7	**99.3**	**50.0**	93.3	**75.0**

Table 4: Comparison of models' results on U-MATH. Scores for various subjects are displayed along with the integral scores. T denotes accuracy over text-only tasks, V denotes accuracy over visual tasks. Asterisk denotes a small number of samples (< 30). Images are not included in the prompt for text-only models, only the problem statements. Note that text-only models can solve a percentage of visual problems, due to either guessing, some of the problems being solvable without the accompanying images, or judgment errors discussed in Section 4.3. **Bold** indicates the best result in each group.

QVQ-72B-Preview: 44.0%), trailing far behind Gemini models. Gemini leads the proprietary category across all scales with considerable margins (Gemini 1.5 Flash: 42.5%, Gemini 1.5 Pro: 47.0%, Gemini 2.0 Flash Thinking: 58.5%).

Visual comprehension is challenging: U-MATH$_{\text{V}}$ scores are consistently much lower compared to U-MATH$_{\text{T}}$, although manual examinations do not suggest the underlying problems to be any harder. Besides, transitioning from text-only to visual often causes degradation in models' textual performance: 48.1% $\Rightarrow$ 42.9% with Mistral and Pixtral Large, 26.1% $\Rightarrow$ 18.6% with smaller Llama-3.1 and Llama-3.2, 71.8% $\Rightarrow$ 59.3% with QwQ and QVQ Preview.

Specialization trumps Size: Larger models expectedly outperform smaller ones, but small-scale specialists like Qwen2.5-Math 7B can surpass models 10 times their size, such as Llama-3.1 70B. Similarly, Qwen2.5-Math 72B performs on par with a 685B mixture-of-experts DeepSeek-V3.

Continuous Finetuning enhances performance: Llama-3.1 70B $\Rightarrow$ Llama-3.1 Nemotron 70B and Qwen2.5-72B $\Rightarrow$ Athene-V2 72B yield 2.9% and 5.2% higher U-MATH accuracy respectively, suggesting that standard-inference models may not be fully optimized for their size and could use high-quality post-training data to improve further.

4.3 Meta-Evaluation (μ-MATH) Results

Meta-evaluations follow the setup in Section 4.1. Additionally, we experiment with two distinct prompting schemes — a standard Automatic Chain-of-Thought (AutoCoT) prompt involving a simple task description followed by an instruction to think step-by-step, and a manual Chain-of-Thought prompt (which we refer to as simply CoT) with explicit instructions on which steps to follow — finding the latter performs best and using it as our default. The judge's output is further processed by an extractor model (Qwen2.5 72B is fixed for consistency), prompted to produce a single label — 'Yes', 'No' or 'Inconclusive' — with 'Inconclusive' reserved for refusals or generation failures and treated as incorrect. Reference Appendix C.2 for the full prompt contents. The main results are presented in Table 5. We summarize our conclusions in the following.

Judgment is non-trivial: In non-reasoners, the maximum attainable F1 score is only 81.5%, and while reasoning models offer significant improvements, reaching a high F1 mark of 90.1%, our results underscore that LLM judges remain fallible — even when applied in an objective domain with access to ground truth labels and using the best current systems. This observation is important because judges' error rates directly limit evaluation precision. Moreover, in cases where judgment errors are systematic in nature as opposed to pure noise — an issue we explore later with an example — this cannot be overcome with sheer data volume.

Judgment is distinct from problem-solving: Superior problem-solving does not necessarily translate to better judgment, as illustrated, for instance, with Qwen2.5 vs. Qwen2.5-Math scores. In fact, our results suggest a trade-off between these skills, tracing to reasoning-coherence tradeoff and manifesting in judges' behavioral differences. These are most apparent (Figure 2) in non-reasoners: proprietary models tend towards conservatism (relatively high TNR compared to TPR), whereas Qwen models, particularly math specialists, exhibit the opposite. See Appendix F for more detailed discussion.

Reasoners exceed the Pareto frontier: Reasoning models improve substantially in both problem-solving and judgment performance over the previous model generation. Notably, the two best performing systems, o1 and o3-mini, are also among the most balanced with respect to TPR-TNR parity.

Prompting effects are substantial yet inhomogeneous across models: In non-reasoners, switching from AutoCoT to CoT generally maintains or improves judgment performance and reduces author bias (see paragraph below), except for Llama models, which suffer an increase in inconclusive judgments (Appendix E, Table 5). Gemini 1.5 models benefit the most (>10% F1 gain), becoming the top non-reasoners and surpassing the Qwen, DeepSeek, and GPT models that beat Gemini with AutoCoT. Reasoner systems, however, remain largely unaffected by the change in prompting.

Judges exhibit model-specific biases: We observe a consistent trend toward better performance on Llama solutions and worse performance on Qwen solutions (see Figure 3). The author bias is most pronounced with smaller judges under AutoCoT prompting and reduced when moving toward more capable models and switching to CoT in the case of non-reasoners. At the same time, no noticeable self-judgment effects are observed.

5 Conclusion

We introduce **U-MATH**, a novel multi-modal benchmark for university-level mathematical reasoning, featuring 1,100 unpublished problems sourced from real teaching materials spanning six university subjects, with 20% involving visual elements. In addition, we provide **μ-MATH**, a U-MATH-derived meta-evaluation dataset enabling rigorous assessment of LLM judges.

Our experiments reveal LLM weaknesses in advanced mathematical reasoning, particularly visual tasks (achieving 58.5% accuracy vs. 93.1% for text-only). Enabling visual reasoning is difficult, often degrading textual performance, and is underdeveloped — especially in open-weight models, which lag significantly behind proprietary ones despite near parity in text-only problems. Nevertheless, continuous fine-tuning, reasoning-first training, and mathematical specialization boost performance, suggesting considerable growth potential.

Judgment proves both distinct from problem-solving and non-trivial for LLMs, with only the most capable models attaining meaningfully high performance while still peaking at an imperfect 90.1% F1-score mark. Additionally, we discover pronounced biases and instabilities in judgment performance as well as distinctive behavioral patterns, underscoring the utility and necessity of meta-evaluations.

Model	U-MATH$_{\text{Text}}$	μ-MATH					μ-MATH$_{\text{Qwen}}$	μ-MATH$_{\text{Llama}}$	μ-MATH$_{\text{GPT}}$	μ-MATH$_{\text{Gemini}}$
		F1$_{\text{CoT}}$ / F1$_{\text{AutoCoT}}$	TPR	TNR	PPV	NPV	F1$_{\text{CoT}}$ / F1$_{\text{AutoCoT}}$	F1$_{\text{CoT}}$ / F1$_{\text{AutoCoT}}$	F1$_{\text{CoT}}$ / F1$_{\text{AutoCoT}}$	F1$_{\text{CoT}}$ / F1$_{\text{AutoCoT}}$
Llama-3.1 8B	33.7	52.0 / 53.1	48.7	55.9	56.0	48.5	48.7 / 49.6	49.2 / 51.2	51.2 / 57.6	55.5 / 50.5
Ministral 8B	26.9	60.5 / 58.9	55.9	65.8	65.4	56.4	52.8 / 55.7	63.1 / 58.2	62.9 / 60.9	58.3 / 54.1
Qwen2.5-Math 7B	53.0	61.9 / 61.2	76.6	47.9	62.9	63.9	59.7 / 56.7	63.8 / 64.0	57.2 / 58.5	63.8 / 61.2
Qwen2.5 7B	50.4	69.3 / 67.0	**78.7**	59.8	69.3	70.8	62.4 / 60.5	72.3 / 72.4	68.3 / 66.4	69.1 / **65.0**
GPT-4o-mini	47.2	72.3 / **69.2**	59.0	88.1	85.1	65.1	69.3 / 61.7	76.2 / **78.5**	**70.4** / **69.8**	69.6 / 64.3
Gemini 1.5 Flash	**61.2**	**74.8** / 65.3	63.3	**88.3**	**86.2**	**67.6**	**71.2** / **61.9**	**80.6** / 70.8	70.1 / 65.3	**73.9** / 59.7
Llama-3.1-70B	40.4	61.0 / 68.2	62.5	59.6	64.1	57.9	56.0 / 63.8	57.0 / 70.2	69.4 / 69.8	58.8 / 64.4
Qwen2.5-Math 72B	68.7	74.0 / 75.5	**80.9**	66.8	73.8	75.2	69.3 / 68.8	77.3 / 79.8	68.2 / 69.2	76.8 / 80.4
Qwen2.5 72B	58.9	75.6 / 75.1	77.1	74.2	77.5	73.7	70.5 / 68.9	79.3 / 80.1	73.7 / 73.4	74.2 / 73.8
Mistral Large	55.6	76.6 / 74.5	75.7	77.7	79.7	73.5	72.5 / 70.8	78.6 / 77.7	76.0 / 74.4	75.0 / 71.0
DeepSeek-V3	69.3	80.6 / **81.5**	77.0	84.7	85.0	76.6	**81.8** / **76.0**	81.2 / **86.2**	74.9 / **80.1**	80.4 / **82.7**
Claude 3.5 Sonnet	40.7	74.8 / 68.1	62.5	**89.5**	**87.3**	67.4	70.8 / 64.1	77.9 / 71.8	72.2 / 68.1	73.8 / 63.4
GPT-4o	53.9	77.4 / 74.2	70.1	85.9	85.1	71.3	74.2 / 68.2	81.8 / 78.9	77.5 / 75.8	72.6 / 70.5
Gemini 1.5 Pro	**71.7**	**81.5** / 69.8	78.5	84.7	85.2	**78.2**	78.9 / 65.4	**83.6** / 74.8	**79.3** / 69.1	**80.5** / 65.8
QwQ-32B-Preview	82.7	81.0 / 79.6	85.7	75.9	80.5	82.2	81.9 / 77.8	81.3 / 79.4	76.1 / 76.8	80.8 / 79.8
DeepSeek-R1	91.3	84.3 / 83.8	77.3	**92.2**	**91.7**	78.4	80.8 / 81.1	87.1 / 85.8	81.8 / 81.5	84.7 / 83.4
Gemini 2.0 Flash-Thinking	89.2	80.2 / 81.2	89.2	70.8	77.4	85.4	77.3 / 78.0	81.1 / 84.0	76.1 / 78.9	82.6 / 79.4
o1-mini	82.9	83.4 / 84.3	78.5	88.8	88.8	78.7	80.0 / 83.8	88.0 / 87.0	81.1 / 82.2	81.3 / 80.8
o3-mini	92.8	89.6 / 89.8	89.0	90.2	91.1	88.0	87.7 / **88.4**	93.2 / 93.6	88.2 / 88.6	86.7 / 85.7
o1	**93.1**	**90.1** / **90.2**	**91.4**	88.6	90.0	**90.2**	**86.1** / 85.7	**94.4** / **94.7**	**88.9** / **89.3**	**88.7** / **89.1**

Table 5: Judgment performance on μ-MATH benchmark using CoT prompting; Macro F1-score (F1), True Positive Rate (TPR), True Negative Rate (TNR), Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are presented, with F1 as the primary metric. The second number within each F1 column written in gray represents the score under AutoCoT prompting. μ-MATH columns display integral scores over the entire benchmark, while μ-MATH $_{\text{<model>}}$ columns denote subsets with solutions generated by specific author models. U-MATH$_{\text{Text}}$ accuracy is added for comparison of each model's performance as a problem-solver vs. as a judge. **Bold** indicates the best result in each column.

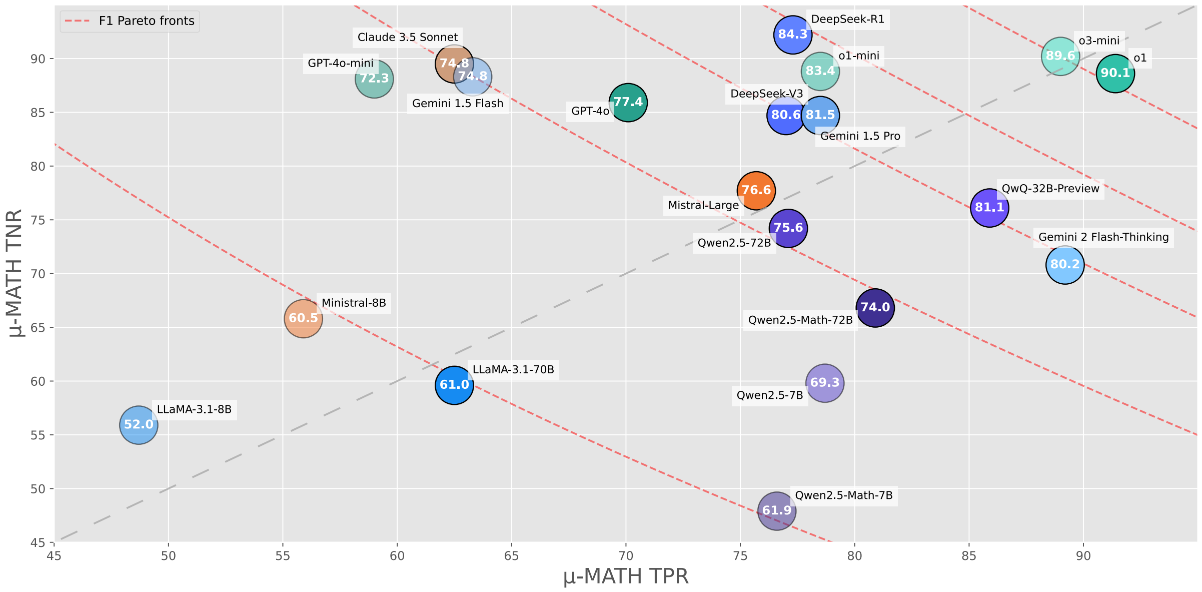


Figure 2: True Positive Rate vs True Negative Rate of judges on μ-MATH. The value inside of the marker denotes the F1-score.

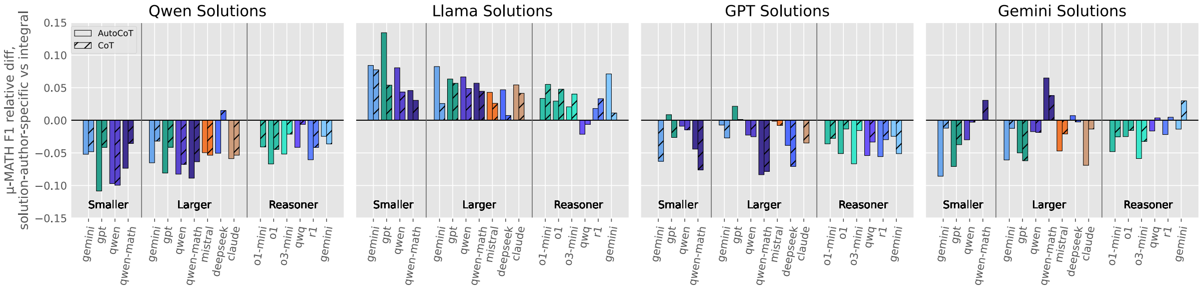


Figure 3: Relative difference in judge μ-MATH F1 scores: performance on a specific author's solutions vs. overall performance. Each pane corresponds to one of the author models. X-axis specifies the judge model (in three groups: small, large, reasoner). Bar pairs compare the difference for AutoCoT vs. manual CoT prompting. The three least performant models (Ministral 8B, Llama-3.1-8B and -70B) are excluded due to outlier behavior (e.g. Appendix E).

Limitations

While U-MATH offers a diverse set of university curricula problems, it does not cover the full range of advanced mathematical subjects. In addition, while the textual parts of our benchmarks demonstrate good model separability across the broad spectrum of recent models, these parts start to approach saturation with the reasoning systems, further necessitating expansion into more advanced topics such as, for example, complex analysis. Moreover, the 20% fraction of visual problems, while reflective of real-world coursework, limits the scope of visual reasoning evaluations. Furthermore, visual problems are not covered by our meta-evaluations.

Although accuracy is a standard metric of choice, it discards a lot of signal and does not allow for finer-grained analyses. Furthermore, reliance on LLM judges introduces errors and biases, and while we do quantify these to some extent, that is only a first step, and additional mitigation mechanisms would need to be put in place in order to account for the errors in a principled manner.

Future Work. Future research can focus on the design of assessment protocols that allow partial credit to enable finer-grained problem-solving evaluations. Another important direction is bridging the gap between quantifying the uncertainty and bias induced by auto-evaluations and controlling for them. Finally, a possible way of overcoming saturation, apart from going through a costly process of curating new data, is coming up with adversarial task creation or modification approaches, which we see as particularly relevant for meta-evaluations. By open-sourcing our data and evaluation code, we strive to facilitate further research and encourage development of models better equipped for complex, real-world mathematical problems.

Ethics Statement

We collected all data in U-MATH and μ-MATH with appropriate permissions, ensuring no personal or proprietary information is included. The datasets consist solely of mathematical problems and solutions, without any sensitive content. We open-sourced the datasets and code under suitable licenses to support transparency and research advancement. There are no known conflicts of interest associated with this work.

Reproducibility Statement

All datasets and evaluation code will be available on GitHub. Detailed descriptions of data collection and processing are presented in Section 3. The experimental setup, including model configurations and prompts, is described in Section 4, with the full prompts provided in Appendices C.1 and C.2.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *Preprint*, arXiv:2402.00157.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthropic. 2024. Introducing claude 3.5 sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`. Accessed: 2024-11-20.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W Ayers, Dragomir Radev, and Jeremy Avigad. 2023. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*.

Edward Beeching, Shengyi Costa Huang, Albert Jiang, Jia Li, Benjamin Lipkin, Zihan Qina, Kashif Rasul, Ziju Shen, Roman Soletskyi, and Lewis Tunstall. 2024. Numinamath 7b cot. `https://huggingface.co/AI-MO/NuminaMath-7B-CoT`.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022a. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3323, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. 2022b. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *Preprint*, arXiv:2105.14517.

Nuo Chen, Ning Wu, Jianhui Chang, and Jia Li. 2024. Controlmath: Controllable data generation promotes math generalist models. *Preprint*, arXiv:2409.15376.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 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.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, 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, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J L Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R J Chen, R L Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S S Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T Wang, Tao Yun, Tian Pei, Tianyu Sun, W L Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X Q Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y K Li, Y Q Wang, Y X Wei, Y X Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z F Wu, Z Z Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. DeepSeek-V3 technical report.

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

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

Google. 2024. gemini2-flash-thinking. `https://de` epmind.google/technologies/gemini/flash-thinking/. Accessed: 2024-10-01.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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. 2022. Solving quantitative reasoning problems with language models. *Preprint*, arXiv:2206.14858.

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. 2024a. Common 7b language models already possess strong math capabilities. *Preprint*, arXiv:2403.04706.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024b. Can multiple-choice questions really be useful in detecting the abilities of llms? *Preprint*, arXiv:2403.17752.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *Preprint*, arXiv:2105.04165.

Yujun Mao, Yoon Kim, and Yilun Zhou. 2024. Champ: A competition-level dataset for fine-grained analyses of llms' mathematical reasoning capabilities. *arXiv preprint arXiv:2401.06961*.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed: 2024-11-15.

Mistral AI. 2024. Announsing pixtral-12b. https://mistral.ai/news/pixtral-12b/. Accessed: 2024-10-01.

Mistral.ai. 2024a. Ministral. https://mistral.ai/en/news/ministraux. Accessed: 2024-10-01.

Mistral.ai. 2024b. Mistral large 2. https://mistral.ai/en/news/mistral-large-2407. Accessed: 2024-10-01.

Mistral.ai. 2024c. Mistral small 3. https://mistral.ai/en/news/mistral-small-3. Accessed: 2024-10-01.

Nexusflow. 2024. Introducing athene-v2: Advancing beyond the limits of scaling with targeted post-training. https://nexusflow.ai/blogs/athene-v2. Accessed: 2024-11-15.

OpenAI. 2024a. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-10-01.

OpenAI. 2024b. o1. https://openai.com/index/learning-to-reason-with-llms/. Accessed: 2024-10-01.

OpenAI. 2024c. o1-mini. https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/. Accessed: 2024-10-01.

OpenAI. 2024d. o3-mini. https://openai.com/index/openai-o3-mini/. Accessed: 2024-10-01.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *Preprint*, arXiv:2308.11483.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. *Preprint*, arXiv:2402.14811.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

QwenLM. 2024a. Qvq 72b preview. https://qwenlm.github.io/blog/qvq-72b-preview/. Accessed: 2024-10-01.

QwenLM. 2024b. Qwq 32b preview. https://qwenlm.github.io/blog/qwq-32b-preview/. Accessed: 2024-10-01.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen, and Benjamin Roth. 2024. From calculation to adjudication: Examining llm judges on mathematical reasoning tasks. *Preprint*, arXiv:2409.04168.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh

Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha

Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024b. Helpsteer2-preference: Complementing ratings with preferences. *Preprint*, arXiv:2410.01257.

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

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2024. Evaluating mathematical reasoning beyond accuracy. *arXiv preprint arXiv:2404.05692*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

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. 2024b. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. 2023. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. *CoRR*, abs/2312.17080.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2024. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2021. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*.

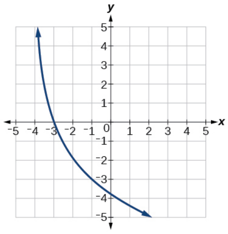
Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Problem examples

A.1 U-MATH Sample Problems

Example 1: Algebra.

Write a logarithmic equation corresponding to the graph, using log base 3:



$$-3\cdot\log_3(x+4)$$

Example 2: Integral Calculus.

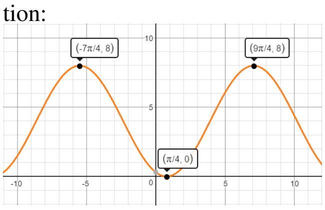
Solve the integral:

$$\int\frac{-9\cdot\sqrt[3]{x}}{9\cdot\sqrt[3]{x^2}+3\cdot\sqrt{x}}\,dx$$

$$-\frac{2}{27}\cdot\ln\left(\frac{|1+3\cdot\sqrt[6]{x}|}{3}\right)-$$
$$-\frac{1}{3}\sqrt[6]{x^2}-\frac{3}{2}\sqrt[6]{x^4}+\frac{2}{3}\sqrt[6]{x^3}+\frac{2}{9}\sqrt[6]{x}+C$$

Example 3: Precalculus Review.

Find a formula for the plotted sinusoidal function:



$$f(x)=-4\cdot\cos\left(\frac{1}{2}\cdot\left(x-\frac{\pi}{4}\right)\right)+4$$

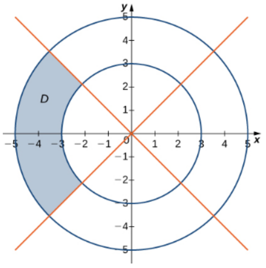
Example 4: Multivariable Calculus.

E is located inside the cylinder $x^2+y^2=1$ and between the circular paraboloids $z=1-x^2-y^2$ and $z=x^2+y^2$. Find the volume of E.

$$\pi/4$$

Example 5: Multivariable Calculus.

The graph of the polar rectangular region D is given. Express the region D in polar coordinates:



1. The interval of r is $[3,5]$
2. The interval of θ is $\left[\frac{3}{4}\cdot\pi,\frac{5}{4}\cdot\pi\right]$

Example 6: Differential Calculus.

Sketch the curve:

$$y=\frac{x^3}{6\cdot(x+3)^2}$$

Provide the following:
1. The domain (in interval notation)
2. Vertical asymptotes
3. Horizontal asymptotes
4. Slant asymptotes
5. Intervals where the function is increasing
6. Intervals where the function is decreasing
7. Intervals where the function is concave up
8. Intervals where the function is concave down
9. Points of inflection

1. The domain: $(-\inf,-3)\cup(-3,\inf)$
...

Figure 4: Examples of text-only and visual problems from the U-MATH benchmark along with their subjects and golden answers.

A.2 U-MATH Sample Problem with Reference Solution

Example: Differential Calculus.

U-MATH Problem:
The function $s(t) = 2 \cdot t^3 - 3 \cdot t^2 - 12 \cdot t + 8$ represents the position of a particle traveling along a horizontal line.
1. Find the velocity and acceleration functions.
2. Determine the time intervals when the object is slowing down or speeding up.

Golden answer:
1. The velocity function $v(t) = 6 \cdot t^2 - 6 \cdot t - 12$ and acceleration function $a(t) = 12 \cdot t - 6$
2. The time intervals when the object speeds up $\left(0, \frac{1}{2}\right)$, $(2, \infty)$ and slows down $\left(\frac{1}{2}, 2\right)$

Reference solution:

The velocity is the derivative of the position function:

$$v(t) = s'(t) = \boxed{6t^2 - 6t - 12}$$

The acceleration is the derivative of the velocity function:

$$a(t) = v'(t) = \boxed{12t - 6}$$

To determine when the object is speeding up or slowing down, we compare the signs of $v(t)$ and $a(t)$.

Step 1: Find the Zeros of $v(t)$ ***and*** $a(t)$
First, solve for $v(t) = 0$:

$$6t^2 - 6t - 12 = 0 \quad \Rightarrow \quad t^2 - t - 2 = 0 \quad \Rightarrow \quad (t-2)(t+1) = 0$$

Thus, $t = 2$ and $t = -1$. Next, solve for $a(t) = 0$:

$$12t - 6 = 0 \quad \Rightarrow \quad t = \frac{1}{2}$$

Step 2: Analyze the Signs of $v(t)$ ***and*** $a(t)$
We analyze the signs of $v(t)$ and $a(t)$ on the intervals determined by $t = -1$, $t = \frac{1}{2}$, and $t = 2$.

Interval	$v(t)$	$a(t)$	Behavior
$(-\infty, -1)$	> 0	< 0	Slowing down
$(-1, \frac{1}{2})$	< 0	< 0	Speeding up
$(\frac{1}{2}, 2)$	< 0	> 0	Slowing down
$(2, \infty)$	> 0	> 0	Speeding up

Step 3: Account for non-negative time

The object is speeding up on $\boxed{\left(0, \frac{1}{2}\right) \text{ and } (2, \infty)}$ and slowing down on $\boxed{\left(\frac{1}{2}, 2\right)}$.

Figure 5: A sample U-MATH problem, including the reference solution and the golden answer.

A.3 μ-MATH Sample Problem

Example: Integral Calculus.

U-MATH Problem:
Solve the integral:

$$\int \frac{20 \cdot \cos(-10 \cdot x)^3}{21 \cdot \sin(-10 \cdot x)^7} \, dx$$

Golden answer:

$$C + \frac{1}{21} \cdot \left(\frac{1}{2} \cdot (\cot(10 \cdot x))^4 + \frac{1}{3} \cdot (\cot(10 \cdot x))^6 \right)$$

LLM-generated answer:

$$-\frac{3\sin(10x)^2 - 2}{126\sin(10x)^6} + C$$

Golden judge verdict: Yes

Comment:
Omitting the arbitrary constants, the reference and the submission could be expressed, respectively, as

$$\frac{\cot^6(10x)}{63} + \frac{\cot^4(10x)}{42} \quad \textit{and} \quad \frac{\csc^6(10x)}{63} - \frac{\csc^4(10x)}{42},$$

which differ by a constant term of $1/126$.

Figure 6: A sample $\boldsymbol{\mu}$-MATH problem, illustrating the comparison between the golden and LLM-generated answers.

B U-MATH Topic Distribution

U-MATH covers a variety of topics across the six of its subjects. Table 6 presents the total number of topics per subject, along with the names and sample counts for the seven most populated topics in each.

Subject	Sample Count	Topic
Differential Calculus	29	Curve Sketching
(51 unique topics)	13	Limits
	12	One-Sided Limits
	12	L'Hospital's Rule
	11	Increasing and Decreasing Functions
	11	Higher Derivatives
	10	Applications of Derivatives (Local Extrema)
Sequences and Series	40	Taylor Series
(28 unique topics)	30	Fourier Series
	18	Maclaurin Series
	12	Approximating Constants Using Power Series
	6	Radius of Convergence (Center of Convergence)
	5	Differentiate Power Series
	4	Error in Approximation
Integral Calculus	83	The Substitution Rule
(35 unique topics)	24	Antiderivatives
	10	Volumes of Solids of Revolution About the X-Axis
	9	Trigonometric Substitutions and Inverse Substitutions
	9	Integrate Respect Independent Variable
	7	Applications of Integrals
	7	Single Variable Surface Area Integrals
Precalculus Review	55	Trigonometric Functions
(19 unique topics)	24	Zeros
	11	Inverses of Functions
	8	Inequalities
	7	Equations with Exponents and Logarithms
	7	Properties of Functions
	6	Exponential Functions
Algebra	18	Equations and Inequalities
(74 unique topics)	13	Polynomial Equations
	8	Find Composition of Two Functions
	7	Polynomials
	6	Find Slope Line
	6	Applications of Exponential Function
	6	Quadratic Equations
Multivariable Calculus	13	Triple Integrals
(53 unique topics)	11	Lagrange Multipliers
	9	Double Integrals in Polar Coordinates
	8	Derivatives of Parametric Equations
	8	Integrals of Multivariable Functions
	8	Double Integral Over General Region
	6	Classification of Critical Points

Table 6: Unique topic counts and top seven populated topics together with their sample sizes per subject.

C Prompts

C.1 Prediction Prompt

Solution CoT Prompt.

`{{problem_statement}}`
Please reason step by step, and put your final answer within \boxed{}

Comment:
Images, if present, are passed by way of a provider-native interface.
For OpenAI-compatible endpoints this is done through the `image_url` *field.*[a]

---

[a] `https://platform.openai.com/docs/guides/vision`

Figure 7: Inference prompt used for sampling solutions given the problem statements.

C.2 Judgment Prompts

Judgment Automatic CoT Prompt.

You’ll be provided with a math problem, a correct answer for it and a solution for evaluation.
You have to answer whether the solution is correct or not.

`---`
PROBLEM STATEMENT:
`{{problem_statement}}`

CORRECT ANSWER:
`{{golden_answer}}`

SOLUTION TO EVALUATE:
`{{model_output}}`
`---`

Now please compare the answer obtained in the solution with the provided correct answer to evaluate whether the solution is correct or not.

Think step-by-step, then conclude with your final verdict by putting either "Yes" or "No" on a separate line.

Figure 8: AutoCoT judgment prompt used for comparing sampled solutions to the golden labels. This prompt variant is only meant for μ-MATH experimentation and has not been used in U-MATH evaluation.

Judgment CoT Prompt.

You’ll be provided with a math problem, a correct answer for it and a solution for evaluation.
You have to answer whether the solution is correct or not.

`---`
PROBLEM STATEMENT:
`{{problem_statement}}`

CORRECT ANSWER:
`{{golden_answer}}`

SOLUTION TO EVALUATE:
`{{model_output}}`
`---`

Now please compare the answer obtained in the solution with the provided correct answer to evaluate whether the solution is correct or not.

Think step-by-step, following these steps, don’t skip any:
1. Extract the answer from the provided solution
2. Make any derivations or transformations that may be necessary to compare the provided correct answer with the extracted answer
3. Perform the comparison
4. Conclude with your final verdict — put either "Yes" or "No" on a separate line

Figure 9: CoT judgment prompt used for comparing sampled solutions to the golden labels. This prompt variant is our default one, and also the one used for U-MATH evaluations.

Judgment Extract Prompt.

You'll be given a result of an evaluation of some mathematical solution by a professional evaluator.
You need to extract the final verdict of this evaluation in simple terms: is the solution graded as correct or not.

Output only a single label — "Yes", "No" or "Inconclusive" — according to the provided evaluation ("Yes" if the solution is graded as correct, "No" if the solution is graded as incorrect, "Inconclusive" if the evaluation is incomplete or the final verdict is not settled upon).

Only output "Inconclusive" for incomplete or unsettled evaluations. If the evaluation does contain a single final verdict like "Yes", "Correct", "True", "No", "Incorrect", "False" and so on, even if it is supplied with some additional disclaimers and remarks, output a "Yes" or "No" label accordingly.

Here goes your input:

```
{{generated_judgment}}
```

Now please output exactly either "Yes", "No" or "Inconclusive".

Figure 10: Prompt for extracting the final verdict from the judge's output.

D U-MATH Visual Comparison

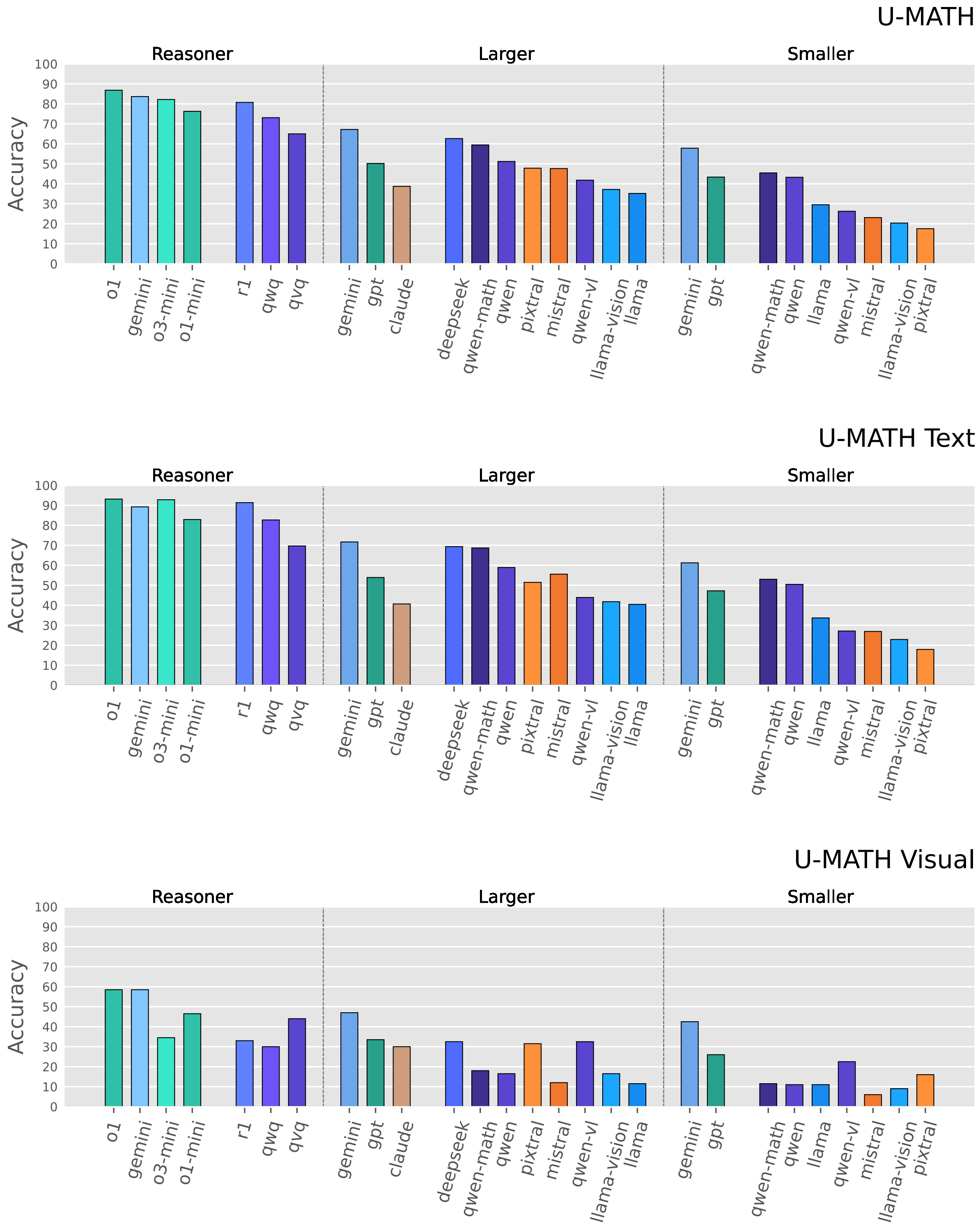


Figure 11: Performance of the selected top-performing models on U-MATH, U-MATH$_{\text{Text}}$ and U-MATH$_{\text{Visual}}$.

E μ-MATH Inconclusive Judgment Rate

Model	**IncRate, AutoCoT**	**IncRate, CoT**
Llama-3.1 8B	13.4	22.9
Llama-3.1 70B	5.0	13.8
Ministral 8B	0.6	5.3
Mistral Large	0.4	1.7
Qwen2.5-Math 7B	2.8	2.4
Qwen2.5-Math 72B	1.2	0.7
Qwen2.5 7B	1.0	1.2
Qwen2.5 72B	1.6	2.1
DeepSeek-V3	0.2	0.2
GPT-4o-mini	0.0	0.1
GPT-4o	0.0	0.0
Gemini 1.5 Flash	0.0	0.1
Gemini 1.5 Pro	0.0	0.0
Claude 3.5 Sonnet	0.0	0.0
QwQ-32B-Preview	0.6	0.9
Gemini 2.0 Flash Thinking	0.2	0.5
DeepSeek-R1	0.0	0.3
o1-mini	0.0	0.1
o1	0.0	0.1
o3-mini	0.0	0.0

Table 7: Percentages of inconclusive judgments produced by each model under different prompting schemes on μ-MATH.

F Problem-solving vs. Judgment, Conservatism vs. Leniency, Reasoning vs. Coherence

This section compares the performance of the models on U-MATH$_{\text{Text}}$ and μ-MATH. The overall score distribution shown in Figure 12 reveals that improved problem-solving capabilities do not necessarily translate to improved judgment. Furthermore, the data suggest **a potential trade-off** between these capabilities, as observed with non-reasoning models, which exhibit a wedge-shaped trend: the two skills improve together up to a certain threshold, beyond which they appear inversely correlated.

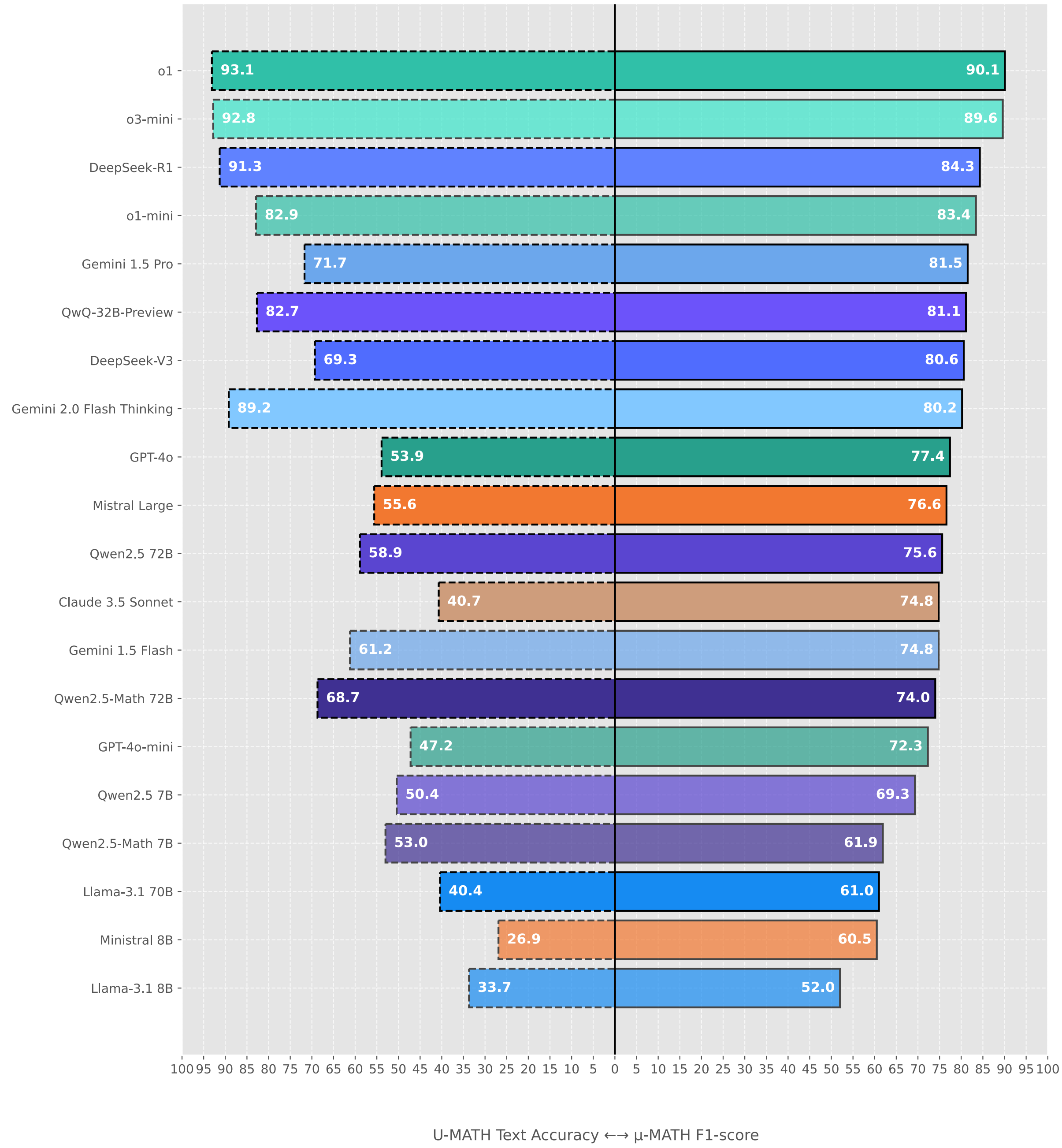


Figure 12: Comparison of LLMs' textual problem-solving (U-MATH$_{\text{Text}}$) vs judgment (μ-MATH) performance.

Based on extensive manual examination, we propose this phenomenon reflects a trade-off between **formal domain-specific reasoning** and **general coherence**. This is perhaps best illustrated by considering the tradeoff's 'extreme ends': Claude Sonnet achieves strong judgment scores despite significantly weaker problem-solving compared to models with similar judgment rankings, something allowing it to compensate for problem-solving deficit, while Qwen-Math, conversely, excels in problem-solving relative to neighbors, indicating some hindrance in translating problem-solving prowess into more effective judgment.

Studying the model responses suggests that what hinders Qwen-Math is exactly the inferior coherence: the model is generally struggling with instruction comprehension, adherence to formatting rules and 'keeping track' of the tasks beyond mathematical problem-solving. Claude, by comparison, is excellent at all of those things, but often to the detriment of in-depth reasoning. To illustrate how this typically plays out, Appendix G provides an example comparing the Claude's and Qwen's judgments on a single μ-MATH sample. Notice how Claude is restrictive and superficial in its comparison, whereas Qwen 'loses the structure' along the way, designating only the first two steps prescribed with the CoT prompt (see prompt contents in Appendix C.2), omitting points three and four and switching to the 'common problem-solving output style'.

We observe this dynamic with all the models to an extent, leading to two corresponding 'judgment styles':

- **Lenient judges:** tend to 'follow the solution', are generally more verbose and good at going into involved derivation chains, which is necessary to arrive at a true positive verdict in more complex scenarios (higher TPR), but comes at a cost of increased hallucination risk and mislabeling negative examples (lower TNR).
- **Conservative judges:** tend to be more 'anchored on the label', are generally more structured and precise, and also less heavy on long hallucination-prone outputs, which reduces the negative mislabeling (higher TNR) but comes at the expense of poor positive recall (lower TPR).

Linking behavioral tendencies to typical outcomes allows us to quantify and visualize these patterns by decomposing the μ-MATH performance into TPR and TNR, as shown in Figure 2. Notice in particular that Claude and Qwen-Math appear as 'the opposites' — having respectively the highest overall TNR and highest overall TPR among the non-reasoners with an approximately equal F1-score.

There are also other patterns emerging, offering deeper insight into the discussed trade-offs.

- **Model tendencies run in the family**: for example, both of the GPT-4 models are conservative, as are both of the Gemini 1.5 models, while all the Qwen models tend to be more lenient. This suggests that these tendencies are largely induced by training data.
- **More balanced training leads to more balanced performance**, as evidenced by comparing the TPR-TNR ratio of Qwen2.5 and Qwen2.5-Math.
- **Losing in capability, on the contrary, exacerbates the bias**, with conservative models mainly losing in TPR and lenient models mainly losing in TNR when moving from a larger model to a smaller one (e.g. Gemini 1.5 Pro ⇒ Gemini 1.5 Flash, GPT-4o ⇒ GPT-4o-mini, Qwen2.5-72B ⇒ Qwen2.5-7B). This indicates that not only a well-balanced training mixture is required but also adequate model capability to generalize over it.
- **Reasoner systems 'push to the right'**, consistent with our observations that increased mathematical problem-solving and verbosity — hallmarks of reasoner systems — correlate with an increase in TPR[2]. This together with previous observations suggests that a potentially promising approach to improving the broad performance and coherence of reasoning systems lies in better training balancing. Notice also that the best performing models, o1 and o3-mini, are among the most balanced ones.

---

[2] Notably, R1 is the only reasoning system that is closer to conservative models in terms of its scores. Upon inspection, we found that its reasoning traces are indeed often driving it towards conservative judgments, the model displaying 'hyper-fixation' over minute details of the golden labels. This is the first case we encountered where an increase in coherence would probably aid more with true positives rather than true negatives. But the sentiment remains the same: coherence and reliability are required to appropriately and successfully apply problem-solving skills to the task at hand.

G μ-MATH Judgment Examples

Example: Differential calculus.

U-MATH Problem:
Compute the derivative of the function

$$y = \sqrt{\frac{x^5 \cdot (2 \cdot x^6 + 3)}{\sqrt[3]{1 - 2 \cdot x}}}$$

Golden answer:

$$y' = \frac{-128 \cdot x^7 + 66 \cdot x^6 - 84 \cdot x + 45}{-24 \cdot x^8 + 12 \cdot x^7 - 36 \cdot x^2 + 18 \cdot x} \cdot \sqrt{\frac{x^5 \cdot (2 \cdot x^6 + 3)}{\sqrt[3]{1 - 2 \cdot x}}}$$

LLM-generated answer:

$$\sqrt{\frac{x^5 \cdot (2x^6 + 3)}{\sqrt[3]{1 - 2x}}} \cdot \frac{1}{2} \left(\frac{5}{x} + \frac{12x^5}{2x^6 + 3} + \frac{2}{3(1 - 2x)} \right)$$

Golden judge verdict: Yes

Claude 3.5 Sonnet Judgment.

1. The solution's answer is: $\sqrt{\frac{x^5 \cdot (2x^6+3)}{\sqrt[3]{1-2x}}} \cdot \frac{1}{2} \left(\frac{5}{x} + \frac{12x^5}{2x^6+3} + \frac{2}{3(1-2x)} \right)$
2. The correct answer is: $\frac{-128 \cdot x^7 + 66 \cdot x^6 - 84 \cdot x + 45}{-24 \cdot x^8 + 12 \cdot x^7 - 36 \cdot x^2 + 18 \cdot x} \cdot \sqrt{\frac{x^5 \cdot \left(2 \cdot x^6 + 3\right)}{\sqrt[3]{1 - 2 \cdot x}}}$
3. To compare these expressions:

- Both have the same factor $\sqrt{\frac{x^5 \cdot \left(2 \cdot x^6 + 3\right)}{\sqrt[3]{1 - 2 \cdot x}}}$
- However, the other factors are completely different:
  - Solution has: $\frac{1}{2} \left(\frac{5}{x} + \frac{12x^5}{2x^6+3} + \frac{2}{3(1-2x)} \right)$
  - Correct answer has: $\frac{-128 \cdot x^7 + 66 \cdot x^6 - 84 \cdot x + 45}{-24 \cdot x^8 + 12 \cdot x^7 - 36 \cdot x^2 + 18 \cdot x}$
- These expressions are not algebraically equivalent, as they have different degrees in numerator and denominator

4. No

Qwen2.5-Math 72B Judgment.

1. The solution provides the derivative as:

$$\frac{dy}{dx} = \sqrt{\frac{x^5 \cdot (2x^6 + 3)}{\sqrt[3]{1 - 2x}}} \cdot \frac{1}{2} \left(\frac{5}{x} + \frac{12x^5}{2x^6 + 3} + \frac{2}{3(1 - 2x)} \right)$$

2. The provided correct answer is:

$$y' = \frac{-128x^7 + 66x^6 - 84x + 45}{-24x^8 + 12x^7 - 36x^2 + 18x} \cdot \sqrt{\frac{x^5 \cdot (2x^6 + 3)}{\sqrt[3]{1 - 2x}}}$$

We need to simplify the expression inside the parentheses of the solution and see if it matches the fraction in the correct answer. First, let's rewrite the expression inside the parentheses:

$$\frac{5}{x} + \frac{12x^5}{2x^6 + 3} + \frac{2}{3(1 - 2x)}$$

To combine these fractions, we need a common denominator. The common denominator is

$$x(2x^6 + 3)(3(1 - 2x)) = 3x(2x^6 + 3)(1 - 2x)$$

Rewriting each term with this common denominator:

$$\frac{5}{x} = \frac{5 \cdot 3(2x^6 + 3)(1 - 2x)}{3x(2x^6 + 3)(1 - 2x)} = \frac{15(2x^6 + 3)(1 - 2x)}{3x(2x^6 + 3)(1 - 2x)}$$

$$\frac{12x^5}{2x^6 + 3} = \frac{12x^5 \cdot 3x(1 - 2x)}{3x(2x^6 + 3)(1 - 2x)} = \frac{36x^6(1 - 2x)}{3x(2x^6 + 3)(1 - 2x)}$$

$$\frac{2}{3(1 - 2x)} = \frac{2 \cdot x(2x^6 + 3)}{3x(2x^6 + 3)(1 - 2x)} = \frac{2x(2x^6 + 3)}{3x(2x^6 + 3)(1 - 2x)}$$

Combining these fractions:

$$\frac{15(2x^6 + 3)(1 - 2x) + 36x^6(1 - 2x) + 2x(2x^6 + 3)}{3x(2x^6 + 3)(1 - 2x)}$$

Expanding the numerator:

$$15(2x^6 + 3)(1 - 2x) = 30x^6 - 60x^7 + 45 - 90x$$

$$2x(2x^6 + 3) = 4x^7 + 6x$$

Adding these together:

$$30x^6 - 60x^7 + 45 - 90x + 36x^6 - 72x^7 + 4x^7 + 6x = -128x^7 + 66x^6 - 84x + 45$$

So the expression inside the parentheses simplifies to:

$$\frac{-128x^7 + 66x^6 - 84x + 45}{3x(2x^6 + 3)(1 - 2x)}$$

Therefore, the derivative is:

$$\sqrt{\frac{x^5 \cdot (2x^6 + 3)}{\sqrt[3]{1 - 2x}}} \cdot \frac{1}{2} \cdot \frac{-128x^7 + 66x^6 - 84x + 45}{3x(2x^6 + 3)(1 - 2x)} = \frac{-128x^7 + 66x^6 - 84x + 45}{6x(2x^6 + 3)(1 - 2x)} \cdot \sqrt{\frac{x^5 \cdot (2x^6 + 3)}{\sqrt[3]{1 - 2x}}}$$

Notice that:

$$-24x^8 + 12x^7 - 36x^2 + 18x = -6x(4x^7 - 2x^6 + 6x - 3) = -6x(2x^6 + 3)(2x - 1) = 6x(2x^6 + 3)(1 - 2x)$$

Thus, the two expressions are identical. Therefore, the solution is $\boxed{\textbf{correct}}$.