ICLR

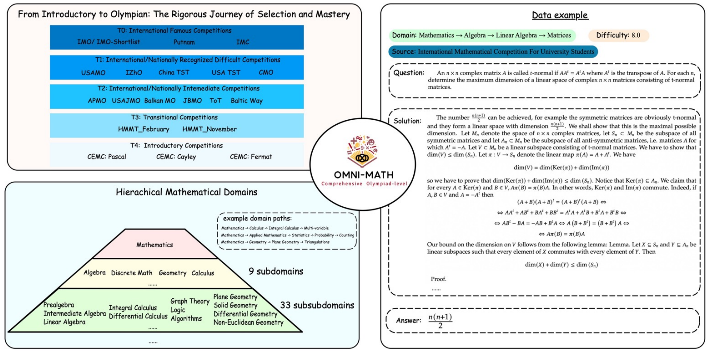
Omni-MATH: A Universal Olympiad Level Mathematic Benchmark for Large Language Models

Bofei Gao*, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu*, Baobao Chang*

北京大学计算机学院
School of Computer Science

通义千问

Introduction


From Introductory to Olympian: The Rigorous Journey of Selection and Mastery
T0: International Famous Competitions
IMO/ IMO-Shortlist
Putnam
IMC
T1: International/Nationally Recognized Difficult Competitions
USAMO
IZhO
China TST
USA TST
CMO
T2: International/Nationally Intemediate Competitions
APMO
USAJMO
Balkan MO
JBMO
ToT
Baltic Way
T3: Transitional Competitions
HMMT_February
HMMT_November
T4: Introductory Competitions
CEMC: Pascal
CEMC: Cayley
CEMC: Fermat
OMNI-MATH
Comprehensive Olympiad-level
Hierachical Mathematical Domains
example domain paths:
Mathematics -> Calculus -> Integral Calculus -> Multi-variable
Mathematics -> Applied Mathematics -> Statistics -> Probability -> Counting
Mathematics -> Geometry -> Plane Geometry -> Triangulations
Mathematics
Algebra
Discrete Math
Geometry
Calculus
......
9 subdomains
Prealgebra
Intermediate Algebra
Linear Algebra
Integral Calculus
Differential Calculus
Graph Theory
Logic
Algorithms
Plane Geometry
Solid Geometry
Differential Geometry
Non-Euclidean Geometry
......
33 subsubdomains
......

Data example

Domain: Mathematics → Algebra → Linear Algebra → Matrices

Difficulty: 8.0

Source: International Mathematical Competition For University Students

Question: An $n \times n$ complex matrix A is called t-normal if $AA^t = A^tA$ where A^t is the transpose of A. For each n, determine the maximum dimension of a linear space of complex $n \times n$ matrices consisting of t-normal matrices.

Solution: The number $\frac{n(n+1)}{2}$ can be achieved, for example the symmetric matrices are obviously t-normal and they form a linear space with dimension $\frac{n(n+1)}{2}$. We shall show that this is the maximal possible dimension. Let M_n denote the space of $n \times n$ complex matrices, let $S_n \subset M_n$ be the subspace of all symmetric matrices and let $A_n \subset M_n$ be the subspace of all anti-symmetric matrices, i.e. matrices A for which $A^t = -A$. Let $V \subset M_n$ be a linear subspace consisting of t-normal matrices. We have to show that $\dim(V) \le \dim(S_n)$. Let $\pi : V \to S_n$ denote the linear map $\pi(A) = A + A^t$. We have

$$\dim(V) = \dim(\mathrm{Ker}(\pi)) + \dim(\mathrm{Im}(\pi))$$

so we have to prove that $\dim(\mathrm{Ker}(\pi)) + \dim(\mathrm{Im}(\pi)) \le \dim(S_n)$. Notice that $\mathrm{Ker}(\pi) \subseteq A_n$. We claim that for every $A \in \mathrm{Ker}(\pi)$ and $B \in V$, $A\pi(B) = \pi(B)A$. In other words, $\mathrm{Ker}(\pi)$ and $\mathrm{Im}(\pi)$ commute. Indeed, if $A, B \in V$ and $A = -A^t$ then

$$(A+B)(A+B)^t = (A+B)^t(A+B) \Leftrightarrow$$
$$\Leftrightarrow AA^t + AB^t + BA^t + BB^t = A^tA + A^tB + B^tA + B^tB \Leftrightarrow$$
$$\Leftrightarrow AB^t - BA = -AB + B^tA \Leftrightarrow A(B + B^t) = (B + B^t)A \Leftrightarrow$$
$$\Leftrightarrow A\pi(B) = \pi(B)A$$

Our bound on the dimension on V follows from the following lemma: Lemma. Let $X \subseteq S_n$ and $Y \subseteq A_n$ be linear subspaces such that every element of X commutes with every element of Y. Then

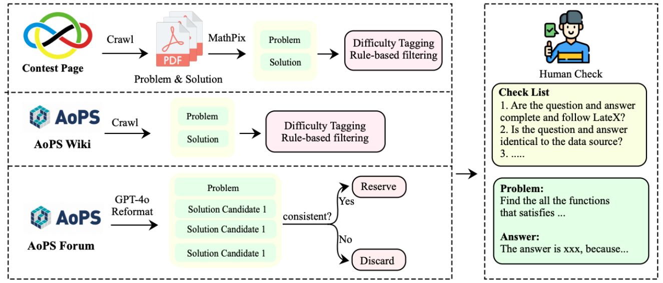
$$\dim(X) + \dim(Y) \le \dim(S_n)$$

Proof.

......

Answer: $\frac{n(n+1)}{2}$

- Omni-MATH is a comprehensive and challenging benchmark specifically designed to assess LLMs' mathematical reasoning at the Olympiad level. Our dataset focuses exclusively on Olympiad mathematics and comprises a vast collection of 4428 competition-level problems. These problems are meticulously categorized into 33 (and potentially more) sub-domains and span across 10 distinct difficulty levels, enabling a nuanced analysis of model performance across various mathematical disciplines and levels of complexity.



Check List
1. Are the question and answer complete and follow LateX?
2. Is the question and answer identical to the data source?
3. .....

Problem:
Find the all the functions that satisfies ...

Answer:
The answer is xxx, because...

Contact Info

- Email: gaobofei@stu.pku.edu.cn
- Paper, code and trained models are available at:
- https://omni-math.github.io

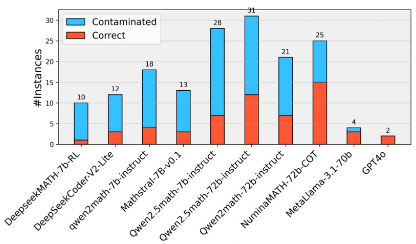
Experiments

Main Result

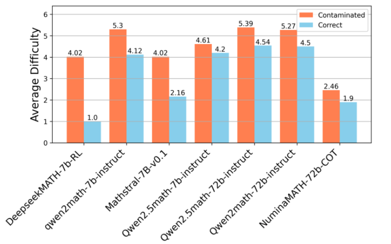
Model	Acc	Alg.	P.Cal	Cal	Geo.	D.M.	Num.	App.	#T1	#T2	#T3	#T4
Vanilla Models												
InternLM2-MATH-mixtral8*22B	14.24	18.19	12.50	10.16	8.70	8.03	10.09	12.36	42.78	8.01	10.35	6.74
DeepSeekMATH-7b-RL	16.12	21.28	20.45	12.50	9.87	7.71	9.98	13.58	49.07	9.11	11.49	7.80
Mathstral-7B-v0.1	19.13	23.99	25.00	13.28	12.19	10.04	14.58	16.30	53.07	10.93	15.29	11.86
DeepSeek-Coder-V2-Lite-Instruct	19.73	24.55	23.86	13.28	13.06	8.92	15.88	16.81	55.93	13.15	12.86	9.55
MetaLlama-3.1-70B-instruct	24.16	29.15	27.59	18.75	14.76	11.74	17.03	24.66	62.66	16.82	16.95	13.71
DeepSeek-Coder-V2	25.78	30.24	35.23	15.62	17.99	12.71	20.90	23.58	65.38	18.84	18.06	14.61
Claude-3.5-SONNET	26.23	30.30	29.55	19.53	17.70	15.74	19.51	26.70	66.23	18.91	18.27	17.41
NuminaMATH-72B-COT	28.45	34.74	27.27	21.88	20.41	16.95	23.47	25.06	65.63	23.70	20.33	21.08
Qwen2-MATH-7b-Instruct	29.36	36.08	35.23	24.22	18.68	14.41	27.04	25.93	63.52	24.30	21.52	18.54
GPT-4o	30.49	36.12	39.77	21.88	21.57	15.74	25.75	29.38	68.38	25.01	21.83	15.81
Qwen2.5-MATH-7b-Instruct	33.22	39.39	37.50	31.25	**26.89**	16.93	28.62	30.37	66.23	29.20	24.68	20.34
Qwen2-MATH-72b-Instruct	33.68	40.27	37.50	27.34	22.53	17.50	30.01	32.96	70.10	29.06	24.71	17.98
Qwen2.5-MATH-72b-Instruct	**36.20**	**43.33**	**42.53**	**39.84**	26.57	**18.28**	**34.28**	**33.37**	**70.96**	**31.37**	**27.75**	**22.29**
Test-time Scaled Models												
Qwen2.5-MATH-7b-Instruct RM@8	35.70	42.12	36.78	33.59	**31.89**	18.96	29.59	30.88	67.95	31.46	27.41	24.0
Qwen2.5-MATH-7b-Instruct RM@256	35.79	42.54	**49.43**	**39.06**	25.79	**19.75**	31.66	33.13	68.24	30.48	**27.81**	23.71
Qwen2.5-MATH-72b-Instruct RM@8	**36.34**	**43.89**	48.28	34.38	26.18	18.28	**33.30**	**34.12**	**71.24**	**32.04**	26.94	23.43
Qwen2.5-MATH-72b-Instruct RM@256	35.95	43.47	47.13	35.94	25.10	19.41	32.64	34.12	68.38	31.46	27.68	**26.28**
OpenAI o1-preview	52.55	57.70	57.47	53.91	43.11	31.26	49.67	53.42	80.11	50.83	42.25	37.71
OpenAI o1-mini	**60.54**	**67.82**	**68.18**	**60.94**	**51.50**	**37.68**	**61.74**	**60.52**	**82.23**	**63.10**	**49.11**	**42.69**

Data Leakage Analysis

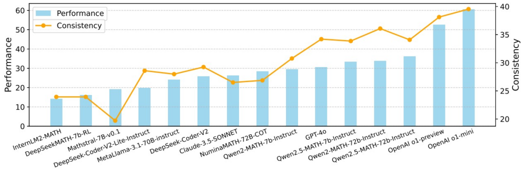
Contaminated and Correct Instances of Different Models



Average difficulty of Contaminated and Correct Instances



Difficulty Consistency Analysis



$$\mathcal{A} = \sum_{i=1}^{n-1} \begin{cases} K \cdot (x_i - x_{i+1}) & \text{if } x_{i+1} > x_i \\ \min(\frac{max(x)}{n}, x_i - x_{i+1}) & \text{if } x_{i+1} <= x_i \end{cases}$$

As the difficulty increases, the overall accuracy of all models declines. Additionally, we note that as model's capability improves, their consistency also increases.

Evaluation

Omni-MATH

Formatting

Problem:
Find all nonnegative integer (x,y,z,w) of the equation\[2^x\cdot3^y-5^z\cdot7^w=1.\]

Reference Answer:
\boxed{(1, 1, 1, 0), (2, 2, 1, 1), (1, 0, 0, 0), (3, 0, 0, 1)}

Model Answer:

```
\boxed{
    \begin{aligned}
    & (1,\ 0,\ 0,\ 0), \\
    & (3,\ 0,\ 0,\ 1), \\
    & (1,\ 1,\ 1,\ 0), \\
    & (2,\ 2,\ 1,\ 1).
\end{aligned}
}
```

Reasoning

Problem:
Find all positive integers $a,n\ge1$ such that..

Reference Answer:
\boxed{(2, 6), (2^k - 1, 2), (1, n) \text{ for any } n \ge 1}

Model Answer:
- a = 1 for any n ≥ 1,
- or n = 2 and a + 1 is a power of 2,
- or a = 2 and n = 6.

Model	LLaMA-2-7b-Chat		LLaMA-3-8b-Instruct		LLaMA-3.1-8b-Instruct	
	Success	Consistency	Success	Consistency	Success	Consistency
MetaLlama-3.1-70B-instruct	98.81	73.63	99.76	77.67	99.76	82.19
DeepSeek-Coder-V2	97.56	38.36	100.00	93.35	100.00	94.01
Qwen2.5-MATH-7b-Instruct	98.45	35.92	99.78	89.80	100.00	90.69
OpenAI o1-preview	99.33	55.03	100.00	90.83	99.78	91.28
OpenAI o1-mini	99.78	63.11	100.00	89.56	100.00	91.78
Mathstral-7B-v0.1	99.33	31.49	99.78	95.12	100.00	95.79
NuminaMATH-72B-COT	100.00	31.11	100.00	88.89	100.00	90.44
Qwen2.5-MATH-72b-Instruct	98.66	37.28	99.78	91.96	100.00	93.30
Total	98.99	45.50	99.89	89.75	99.94	91.26

Our experiments reveal that **GPT-4o evaluation can align well with human evaluation with an accuracy of 98%** and the Omni-Judge achieves over 90% consistency with GPT-4o, providing an efficient and reliable evaluation.

Rule-based Evaluation Subset

Model	Acc @Rule 2821	Acc @GPT-4o 4428
o1-mini	62.2%	60.54%
o1-preview	51.7%	52.55%
qwen2.5-MATH-72b-Instruct	35.7%	36.20%
qwen2.5-MATH-7b-Instruct	32.3%	33.22%
GPT-4o	29.2%	30.49%
NuminaMATH-72b-cot	27.1%	28.45%
DeepseekMATH-7b-RL	14.9%	16.12%

Latest News 🔥

- Omni-MATH is cited by many famous work like: *Kimi-k1.5, Qwen-MATH, Frontier MATH, Livebench, S1, Lmm-r1.*
- Omni-MATH is adopted for training by many famous project like: *Big-Math-Verified project, DeepScaleR, DeepMath-103K, Meta-CoT.*