OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems

Chaoqun He, et al

Tsinghua University

June 6, 2024

Abstract

- This benchmark consists of problems from exam, competitions
- Each problem is detailed with expert-level annotations
- Even GPT-4V attained an average score of 17.97%

Introduction

Question: Find all triples (x, y, z) of positive integers such that $x \le y \le z$ and $x^3(y^3 + z^3) = 2012(xyz + 2)$.

Solution: First note that x divides $2012 \cdot 2 = 2^3 \cdot 503$. If $503 \mid x$ then the right-hand side of the equation is divisible by 503^3, and it follows that $503^2 \mid xyz + 2$. This is false as $503 \mid x$. Hence $x = 2^m$ with $m \in \{0,1,2,3\}$. If $m \ge 2$ then $2^6 \mid 2012(xyz + 2)$. However the highest powers of 2 dividing 2012 and $xyz + 2 = 2^m yz + 2$ are 2^2 and 2^1 respectively. So $x = 1$ or $x = 2$, yielding the two equations

$$y^3 + z^3 = 2012(yz + 2),$$
$$y^3 + z^3 = 503(yz + 1)$$

In both cases …… It follows that $y \equiv -z (mod\ 503)$ as claimed. Therefore $y + z = 503k$ with $k \ge 1$. In view of $y^3 + z^3 = (y + z)((y - z)^2 + yz)$ the two equations take the form

$$k(y - z)^2 + (k - 4)yz = 8 \quad (1)$$
$$k(y - z)^2 + (k - 1)yz = 1 \quad (2)$$

In (1) we have $(k - 4)yz \le 8$, which implies k ≤ 4 ……
Therefore (1) has no integer solutions.
Equation (2) implies $0 \le (k - 1)yz \le 1$, so that $k = 1$ or $k = 2$.
Also $0 \le k(y - z)^2 \le 1$, hence $k = 2$ only if $y = z$. However then $y = z = 1$, which is false in view of $y + z \ge 503$.
Therefore $k = 1$ and (2) takes the form $(y - z)^2 = 1$, yielding $z - y = |y - z| = 1$. Combined with $k = 1$ and $y + z = 503k$, this leads to $y = 251, z = 252$.
In summary the triple (2,251,252) is the only solution.

- Many benchmarks lack sufficient challenge for the latest models
  - GPT-4 with prompting techniques has achieved 97.0% on GSM8K
  - They focus on text, unable to understand geometry and physics

Dataset

Benchmark	Subject		Multi-	Detailed	Difficulty	Size		Answer	Language	Question
	Maths	Physics	modal	solution	level	Maths	Physics	type	type	type
SciBench	✓	✓	✓	✓	COL	217	295	Num	EN	OE
MMMU	✓	✓	✓	✓	COL	540	443	Num	EN	MC,OE
MathVista	✓		✓		-	1,000		Num	EN	MC,OE
ScienceQA		✓	✓		H		617		EN	MC
SciEval		✓			-		1,657	Num	EN	MC,FB,J
JEEBench	✓	✓		✓	CEE	236	123	Num	EN	MC,OE
MMLU	✓	✓			COL	948	548		EN	MC
AGIEval	✓	✓			CEE	953	200	Num	EN,ZH	MC,FB,OE
GSM8K	✓			✓	E	1,319		Num	EN	OE
MATH	✓			✓	COMP	5,000		Num,Exp,Tup	EN	OE
OlympiadBench	✓	✓	✓	✓	COMP	6,142	2,334	ALL	EN,ZH	OE

- Difficulty level (COMP, COL, CEE, H, E)

  - Competition, College, College Entrance Exam, High / Elementary School

- Bilingual

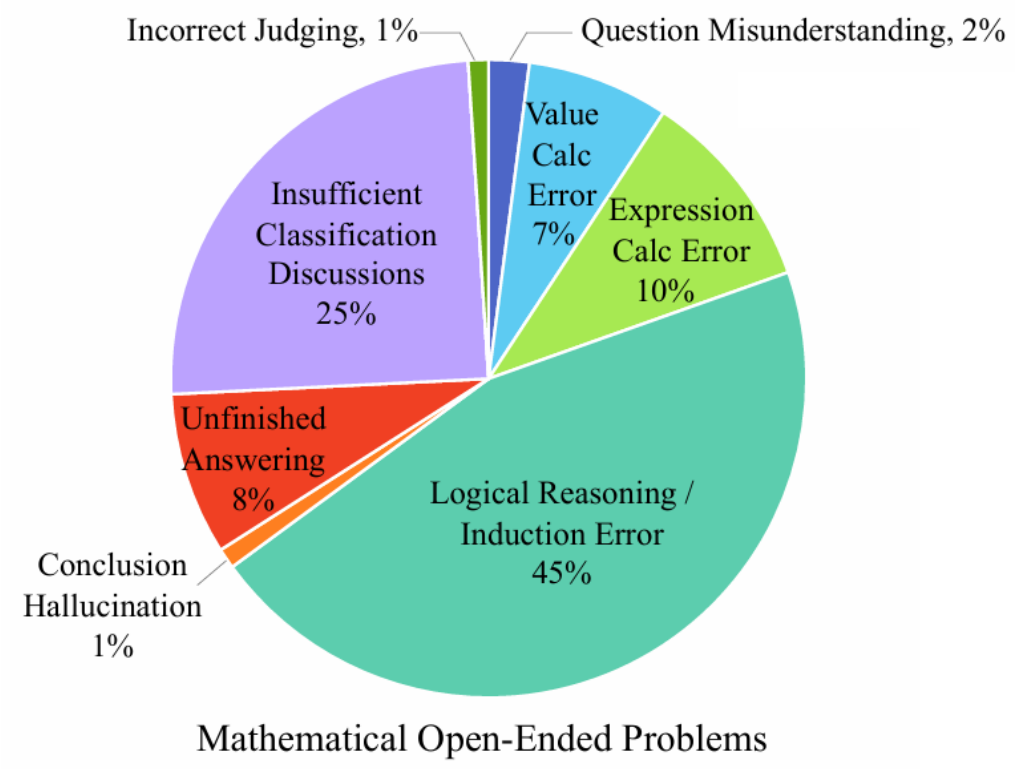
  - EN for English, ZH for Chinese

Experiment

Models	Maths					Physics			Avg.
	En_COMP	Zh_COMP	Zh_CEE	Avg.		En_COMP	Zh_CEE	Avg.	
LLaVA-NeXT-34B†	3.98	2.60	4.64	4.30	-	1.36	2.32	2.08	3.65
Yi-VL-34B†	4.22	3.68	4.31	4.23	-	0.91	1.64	1.46	3.42
Gemini-Pro-Vision	6.92	2.59	5.05*	5.14	-	3.19*	2.12	2.45	4.22
Qwen-VL-Max	10.68	13.21*	13.08	12.65	-	3.76*	5.64*	5.09	10.09
GPT-4V	27.18	14.87	21.27	21.70	-	11.42	10.45	10.74	17.97
Experiment with text-only									
LLaVA-NeXT-34B	4.15	2.94	8.55	6.29	-	2.12	5.22	3.13	5.87
Yi-VL-34B	4.45	3.68	8.06	6.24	-	0.85	5.22	2.28	5.72
DeepSeekMath-7B-RL	19.44	2.70	22.42	18.09	-	6.78	16.52	9.97	17.02
Gemini-Pro-Vision	7.57	2.94	9.20*	7.63	-	4.66	6.96	5.41	7.34
Qwen-VL-Max	11.57	14.29	25.89	19.70	-	4.24	18.26	8.83	18.27
GPT-4V	28.93	15.93	37.10	31.01	-	12.71	23.48	16.24	29.07
GPT-4	30.42	16.42	37.98	32.00	-	12.29	24.35	16.24	29.93

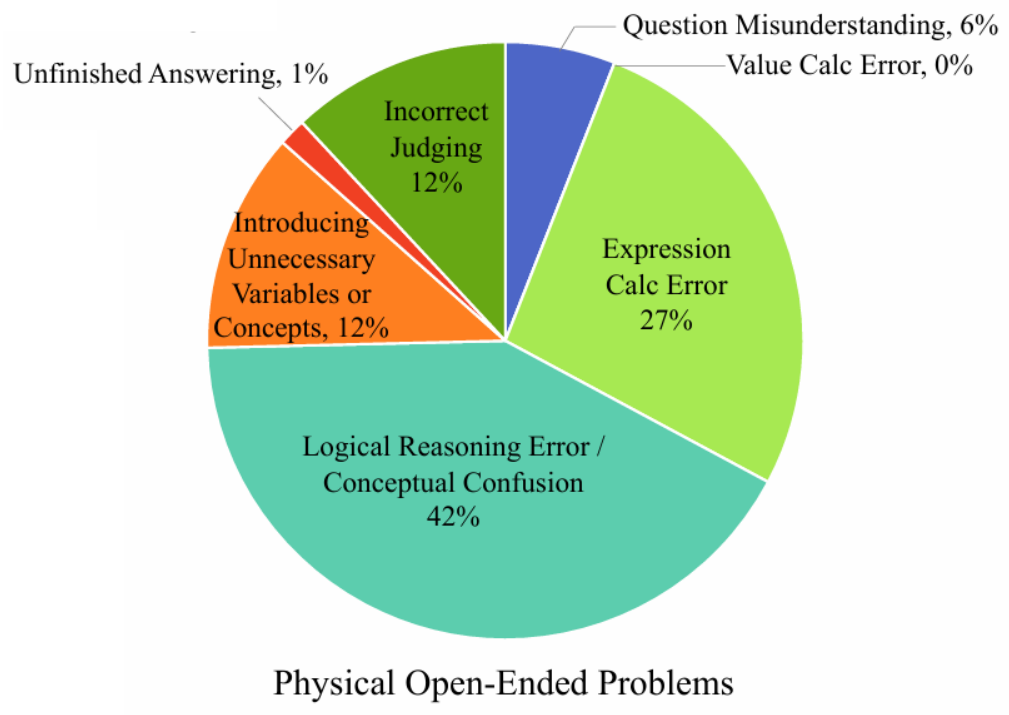
- Performance drop in physics, Chinese, image problems

Experiment



- Mathematics Error
  - Logical Reasoning Error(45%)
  - Insufficient classification in Combinatorial problems(25%)
  - Large Calculation Error(10%)

Experiment



- Physics Error
  - Logical Reasoning Error(42%)
  - Large Calculation Error(25%)
  - Introducing unnecessary variables(12%)

Conclusion

- We proposed advanced benchmark
  - Each problem is detailed with expert-level annotations
  - Pinpointing prevalent error types
- Limitations
  - There are proof problems
  - Code generation or automated verification is impossible