NEURAL INFORMATION PROCESSING SYSTEMS

OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI

Zhen Huang[3,4], Zengzhi Wang[1,4], Shijie Xia[1,4], Xuefeng Li[1,4], Haoyang Zou[4], Ruijie Xu[1,4], Run-Ze Fan[1,4], Lyumanshan Ye[1,4], Ethan Chern[1,4], Yixin Ye[1,4], Yikai Zhang[1,4] Yuqing Yang[4], Ting Wu[4], Binjie Wang[4], Shichao Sun[4], Yang Xiao[4], Yiyuan Li[4], Fan Zhou[1,4] Steffi Chern[4], Yiwei Qin[4], Yan Ma[4], Jiadi Su[4], Yixiu Liu[1,4], Yuxiang Zheng[1,4] Shaoting Zhang[2], Dahua Lin[2], Yu Qiao[2], Pengfei Liu[1,2,4]

[1]Shanghai Jiao Tong University, [2]Shanghai Artificial Intelligence Laboratory,
[3]Soochow University, [4]Generative AI Research Lab (GAIR)

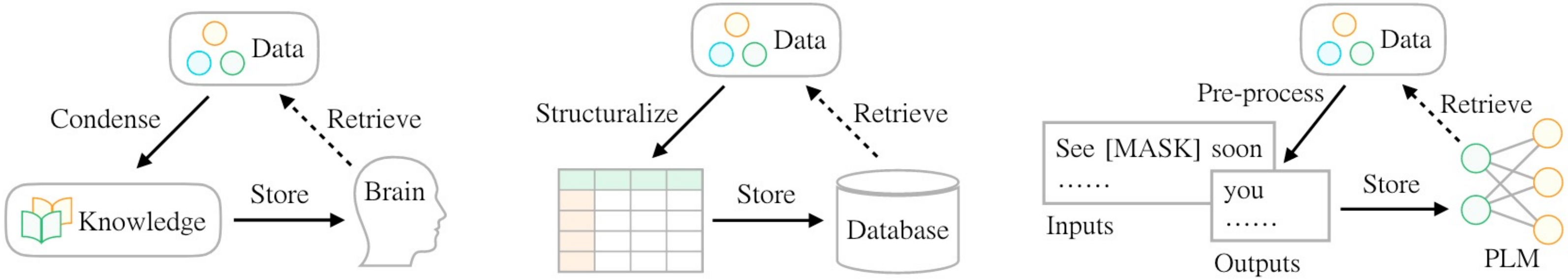
Background

How to benchmark AI Intelligence?

Stage1: Focus on specialized domains (CV: MNIST, ImageNet, NLP: GLUE, XTREME).

The success of LLMs.

Pre-train, Prompt, and Predict



(a) Biological neural networks.

(b) Disk/Cloud storage.

(c) Artificial neural netwokrs.

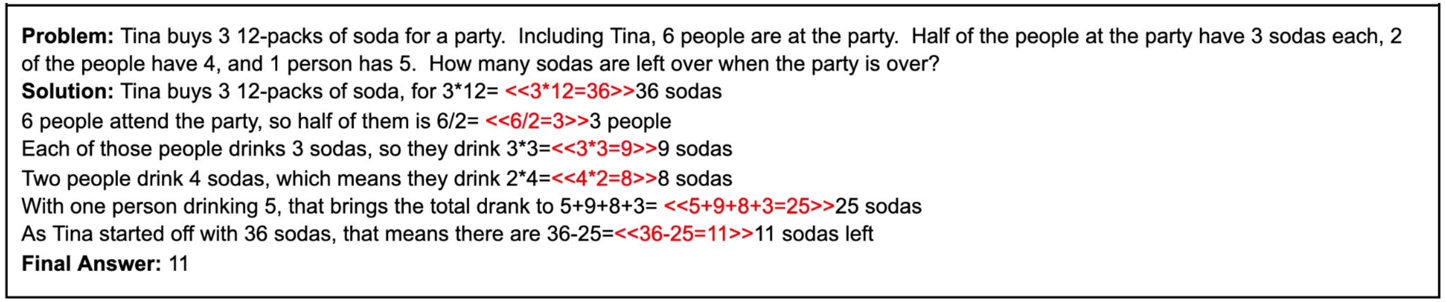
Stage2: Emphasize the evaluation of foundational knowledge and innate abilities (MMLU, C-Eval).

LLMs are quite good at these knowledge-intensive tasks.

Stage3: ? AGI (Artificial General Intelligence) → Superintelligence

Related Works

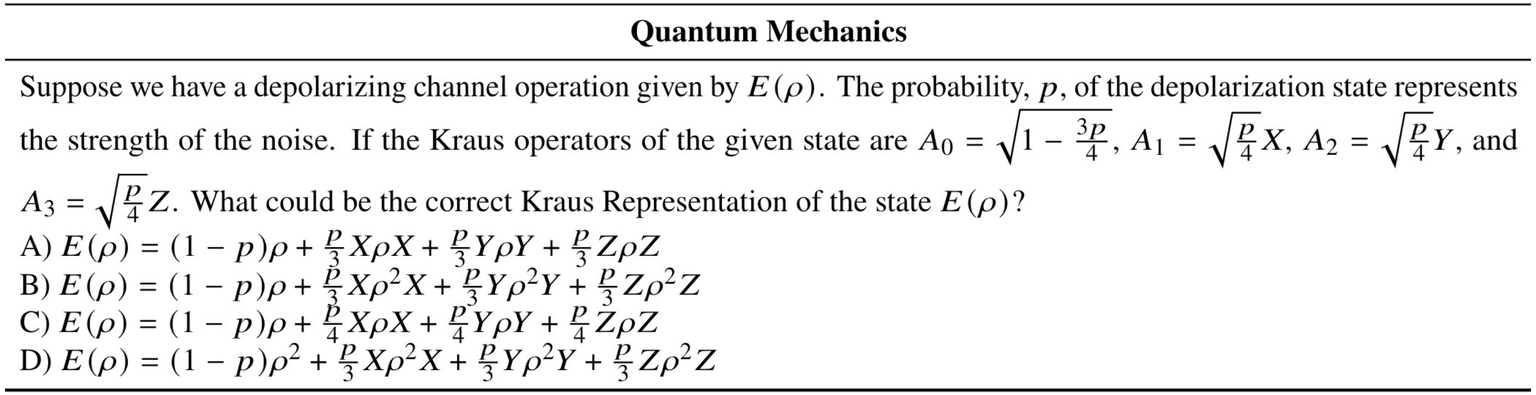
Direction 1: From knowledge-intensive tasks to reasoning-intensive tasks.



Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?
Solution: Tina buys 3 12-packs of soda, for 3*12= <<3*12=36>>36 sodas
6 people attend the party, so half of them is 6/2= <<6/2=3>>3 people
Each of those people drinks 3 sodas, so they drink 3*3=<<3*3=9>>9 sodas
Two people drink 4 sodas, which means they drink 2*4=<<4*2=8>>8 sodas
With one person drinking 5, that brings the total drank to 5+9+8+3= <<5+9+8+3=25>>25 sodas
As Tina started off with 36 sodas, that means there are 36-25=<<36-25=11>>11 sodas left
Final Answer: 11

GSM-8K, MATH

Direction 2: From single discipline (i.e. Math) to multi-discipline.



Quantum Mechanics

Suppose we have a depolarizing channel operation given by $E(\rho)$. The probability, p, of the depolarization state represents the strength of the noise. If the Kraus operators of the given state are $A_0 = \sqrt{1 - \frac{3p}{4}}$, $A_1 = \sqrt{\frac{p}{4}}X$, $A_2 = \sqrt{\frac{p}{4}}Y$, and $A_3 = \sqrt{\frac{p}{4}}Z$. What could be the correct Kraus Representation of the state $E(\rho)$?
A) $E(\rho) = (1-p)\rho + \frac{p}{3}X\rho X + \frac{p}{3}Y\rho Y + \frac{p}{3}Z\rho Z$
B) $E(\rho) = (1-p)\rho + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$
C) $E(\rho) = (1-p)\rho + \frac{p}{4}X\rho X + \frac{p}{4}Y\rho Y + \frac{p}{4}Z\rho Z$
D) $E(\rho) = (1-p)\rho^2 + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$

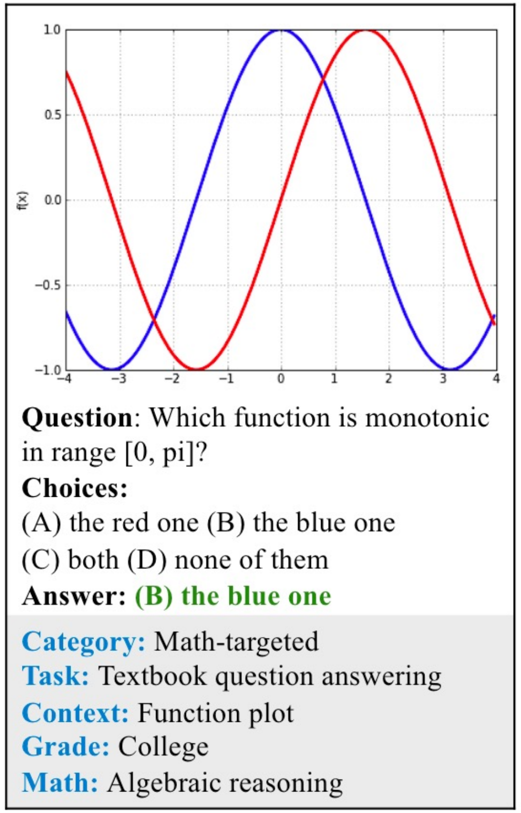
GPQA: Graduate-level multiple-choice questions

Related Works

NEURAL INFORMATION PROCESSING SYSTEMS

Direction 3: From text-only to multi-modal.

Human cognition integrates multiple sensory inputs such as visual information.



Question: Which function is monotonic in range [0, pi]?

Choices:

(A) the red one (B) the blue one

(C) both (D) none of them

Answer: (B) the blue one

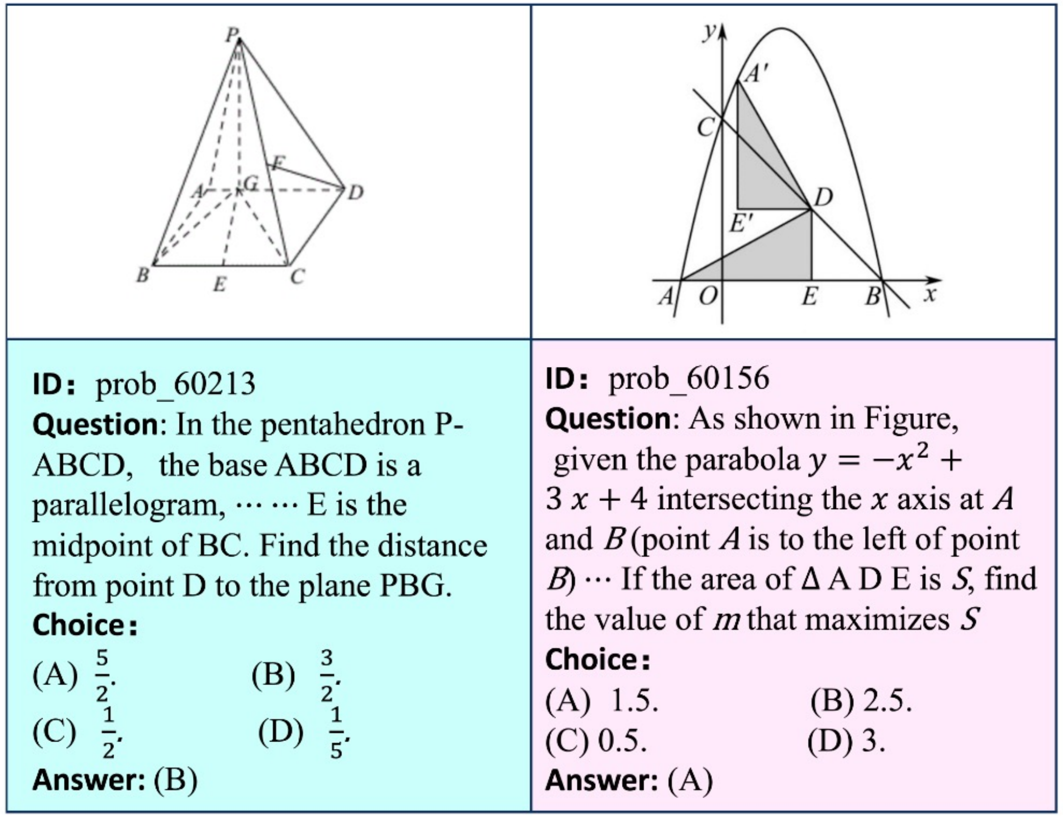
Category: Math-targeted

Task: Textbook question answering

Context: Function plot

Grade: College

Math: Algebraic reasoning



ID: prob_60213

Question: In the pentahedron P-ABCD, the base ABCD is a parallelogram, ⋯ ⋯ E is the midpoint of BC. Find the distance from point D to the plane PBG.

Choice:

(A) $\frac{5}{2}$. (B) $\frac{3}{2}$.

(C) $\frac{1}{2}$. (D) $\frac{1}{5}$.

Answer: (B)

ID: prob_60156

Question: As shown in Figure, given the parabola $y = -x^2 + 3\,x + 4$ intersecting the x axis at A and B (point A is to the left of point B) ⋯ If the area of Δ A D E is S, find the value of m that maximizes S

Choice:

(A) 1.5. (B) 2.5.

(C) 0.5. (D) 3.

Answer: (A)

Limitations

Limitations of exsiting scientific problem-solving benchmarks and how we solve:

□ The challenge is not sufficient, it no longer poses a difficulty for current LLMs.

Dataset	Type	Accuracy of GPT-4o
GSM8K	Grade School	92.0
MATH	High School	76.6

OpenAI O1 achieves 94.8% acc on MATH.

Olympic-level problems are suitable !

Limitations

Limitations of exsiting scientific problem-solving benchmarks and how we solve:

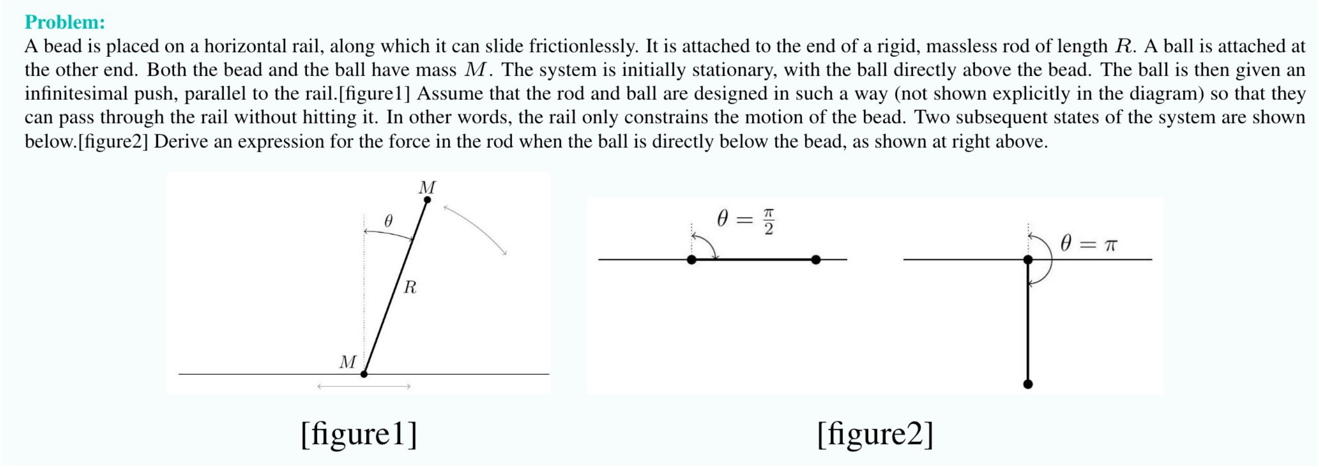
□ The challenge is not sufficient, it no longer poses a difficulty for current LLMs.

Dataset	Type	Accuracy of GPT-4o
GSM8K	Grade School	92.0
MATH	High School	76.6

OpenAI O1 achieves 94.8% acc on MATH.

Olympic-level problems are suitable !

□ Lack a **comprehensive** benchmark that is reasoning-intensive, multi-discipline, and multi-modal.



Problem:

A bead is placed on a horizontal rail, along which it can slide frictionlessly. It is attached to the end of a rigid, massless rod of length R. A ball is attached at the other end. Both the bead and the ball have mass M. The system is initially stationary, with the ball directly above the bead. The ball is then given an infinitesimal push, parallel to the rail.[figure1] Assume that the rod and ball are designed in such a way (not shown explicitly in the diagram) so that they can pass through the rail without hitting it. In other words, the rail only constrains the motion of the bead. Two subsequent states of the system are shown below.[figure2] Derive an expression for the force in the rod when the ball is directly below the bead, as shown at right above.

[figure1]

[figure2]

7 disciplines: Math, Physics, Chemistry, Biology, Geography, Astronomy, CS

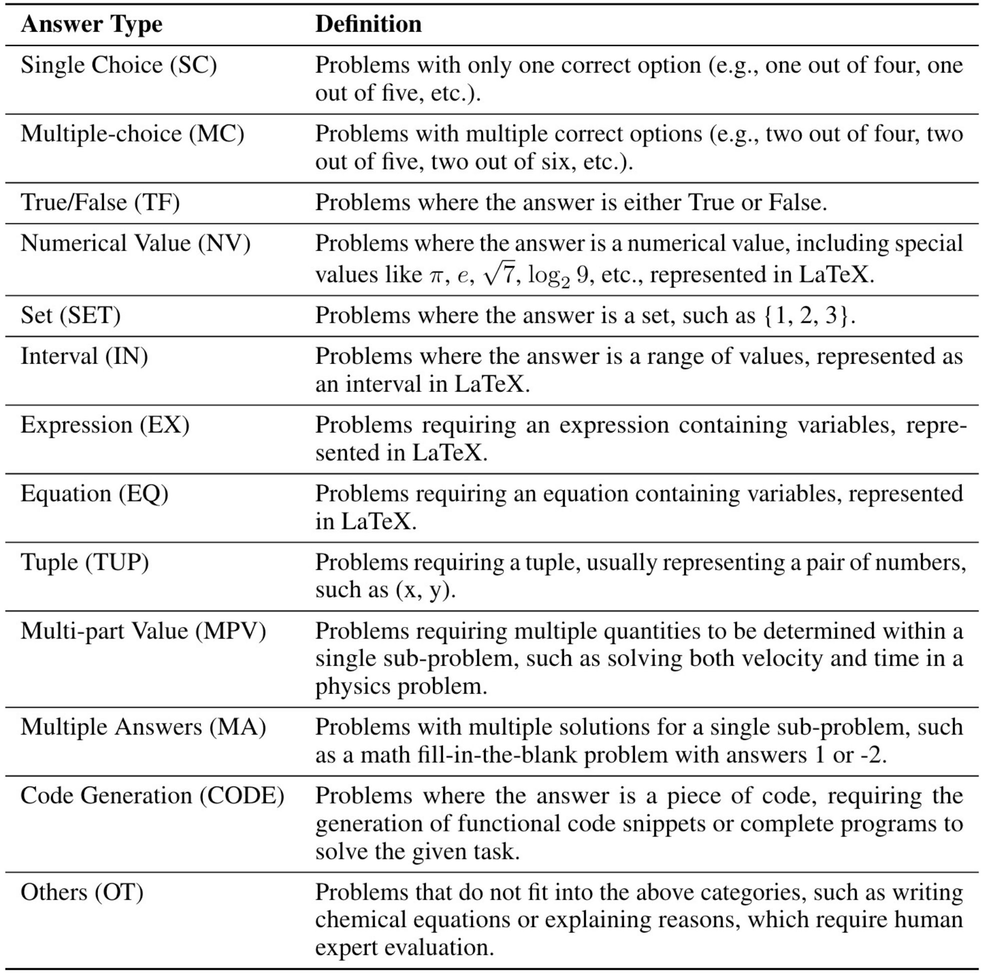
62 different competitions, 34 branches

Single image -> **Interleaved** image-text inputs

Limitations

Limitations of exsiting scientific problem-solving benchmarks and how we solve:

□ Limited to only a few objective question types (such as multiple-choice, true/false, and fill-in-the-blank).



Answer Type	Definition
Single Choice (SC)	Problems with only one correct option (e.g., one out of four, one out of five, etc.).
Multiple-choice (MC)	Problems with multiple correct options (e.g., two out of four, two out of five, two out of six, etc.).
True/False (TF)	Problems where the answer is either True or False.
Numerical Value (NV)	Problems where the answer is a numerical value, including special values like π, e, $\sqrt{7}$, $\log_2 9$, etc., represented in LaTeX.
Set (SET)	Problems where the answer is a set, such as $\{1, 2, 3\}$.
Interval (IN)	Problems where the answer is a range of values, represented as an interval in LaTeX.
Expression (EX)	Problems requiring an expression containing variables, represented in LaTeX.
Equation (EQ)	Problems requiring an equation containing variables, represented in LaTeX.
Tuple (TUP)	Problems requiring a tuple, usually representing a pair of numbers, such as (x, y).
Multi-part Value (MPV)	Problems requiring multiple quantities to be determined within a single sub-problem, such as solving both velocity and time in a physics problem.
Multiple Answers (MA)	Problems with multiple solutions for a single sub-problem, such as a math fill-in-the-blank problem with answers 1 or -2.
Code Generation (CODE)	Problems where the answer is a piece of code, requiring the generation of functional code snippets or complete programs to solve the given task.
Others (OT)	Problems that do not fit into the above categories, such as writing chemical equations or explaining reasons, which require human expert evaluation.

Rule-based Evaluation

Model-based Evaluation (with meta-evaluation)

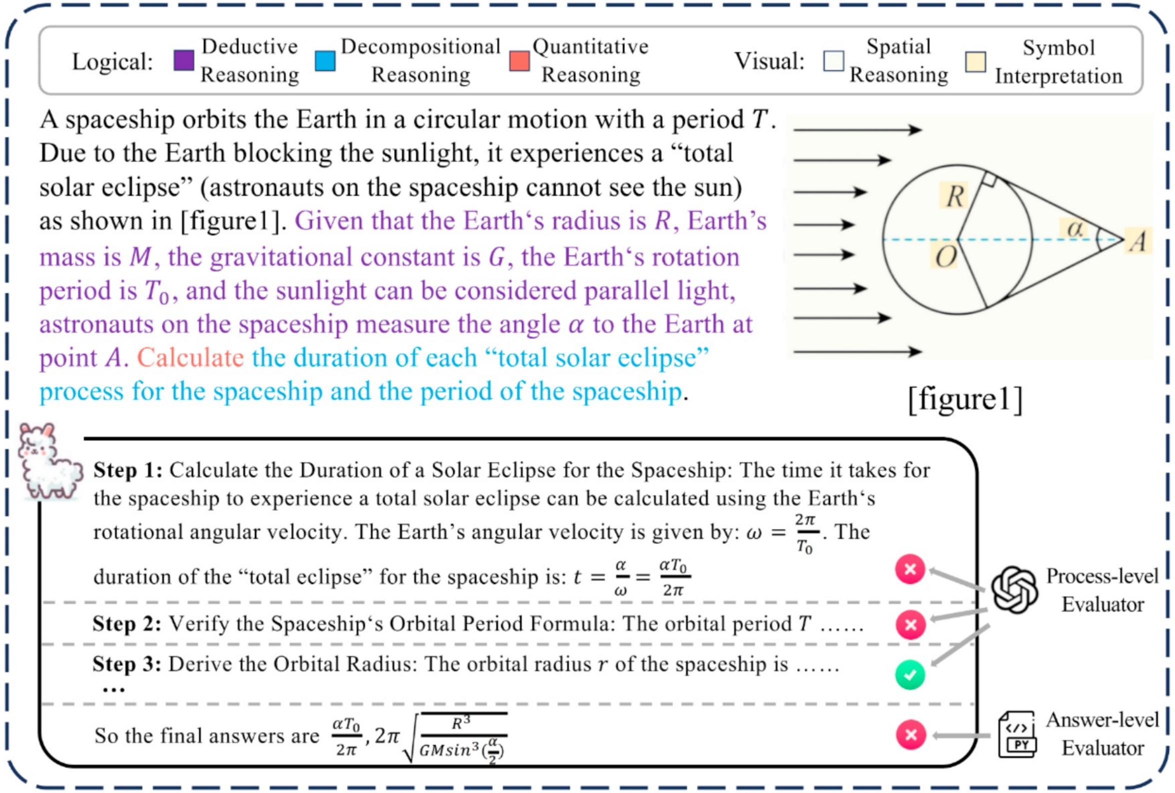
13 different answer types

Limitations

NEURAL INFORMATION PROCESSING SYSTEMS

Limitations of exsiting scientific problem-solving benchmarks and how we solve:

- □ Existing benchmarks often focus solely on answer-level evaluation, lacking **process-level evaluation**.
- □ Existing evaluations lack assessments of different **fine-grained reasoning abilities**.



Logical: 🟪 Deductive Reasoning 🟦 Decompositional Reasoning 🟥 Quantitative Reasoning Visual: □ Spatial Reasoning 🟨 Symbol Interpretation

A spaceship orbits the Earth in a circular motion with a period T. Due to the Earth blocking the sunlight, it experiences a "total solar eclipse" (astronauts on the spaceship cannot see the sun) as shown in [figure1]. Given that the Earth's radius is R, Earth's mass is M, the gravitational constant is G, the Earth's rotation period is T_0, and the sunlight can be considered parallel light, astronauts on the spaceship measure the angle α to the Earth at point A. Calculate the duration of each "total solar eclipse" process for the spaceship and the period of the spaceship.

[figure1]

Step 1: Calculate the Duration of a Solar Eclipse for the Spaceship: The time it takes for the spaceship to experience a total solar eclipse can be calculated using the Earth's rotational angular velocity. The Earth's angular velocity is given by: $\omega = \frac{2\pi}{T_0}$. The duration of the "total eclipse" for the spaceship is: $t = \frac{\alpha}{\omega} = \frac{\alpha T_0}{2\pi}$ ✗

Step 2: Verify the Spaceship's Orbital Period Formula: The orbital period T …… ✗

Step 3: Derive the Orbital Radius: The orbital radius r of the spaceship is …… ✓

…

So the final answers are $\frac{\alpha T_0}{2\pi}, 2\pi\sqrt{\frac{R^3}{GMsin^3(\frac{\alpha}{2})}}$ ✗

You are a teacher skilled in evaluating the intermediate steps of a student's solution to a given problem.

You are given two types of step-by-step solutions: one from the reference answer and the other from the student. Your task is to evaluate the correctness of each step in the student's solutions using binary scoring: assign a score of 1 for correct steps and 0 for incorrect steps. Use the reference solutions to guide your evaluation.
Follow the format:
Step 1: ...
Step 2: ...
Step 3: ...
Please provide the results directly, omitting any introductory or concluding remarks.
The given question

{the question}

The reference solution

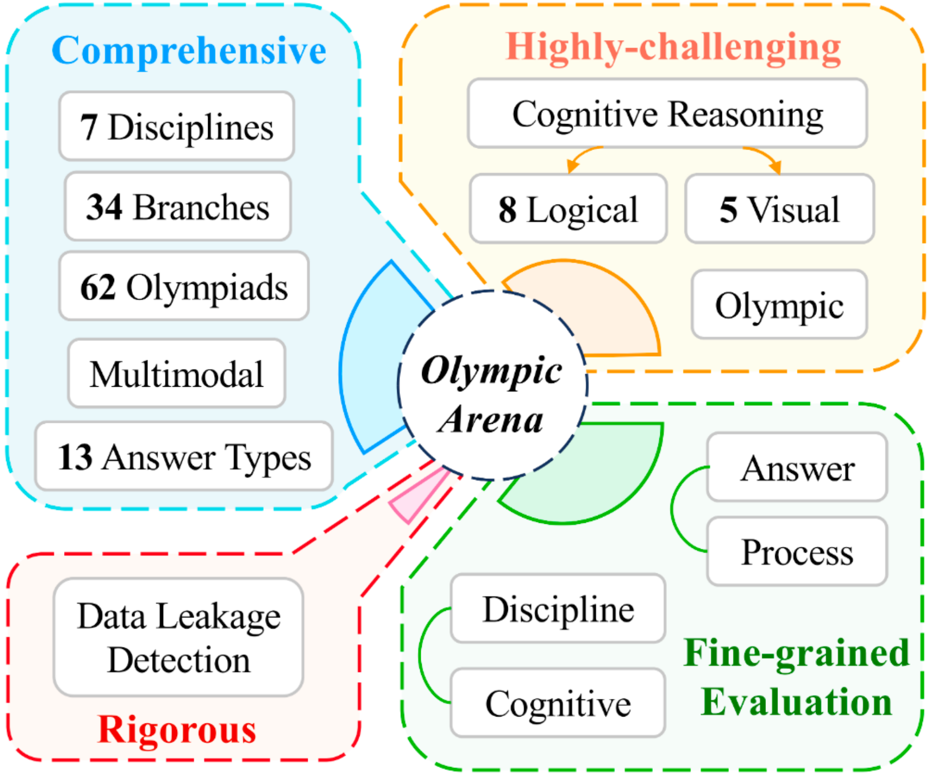
{the reference solution}

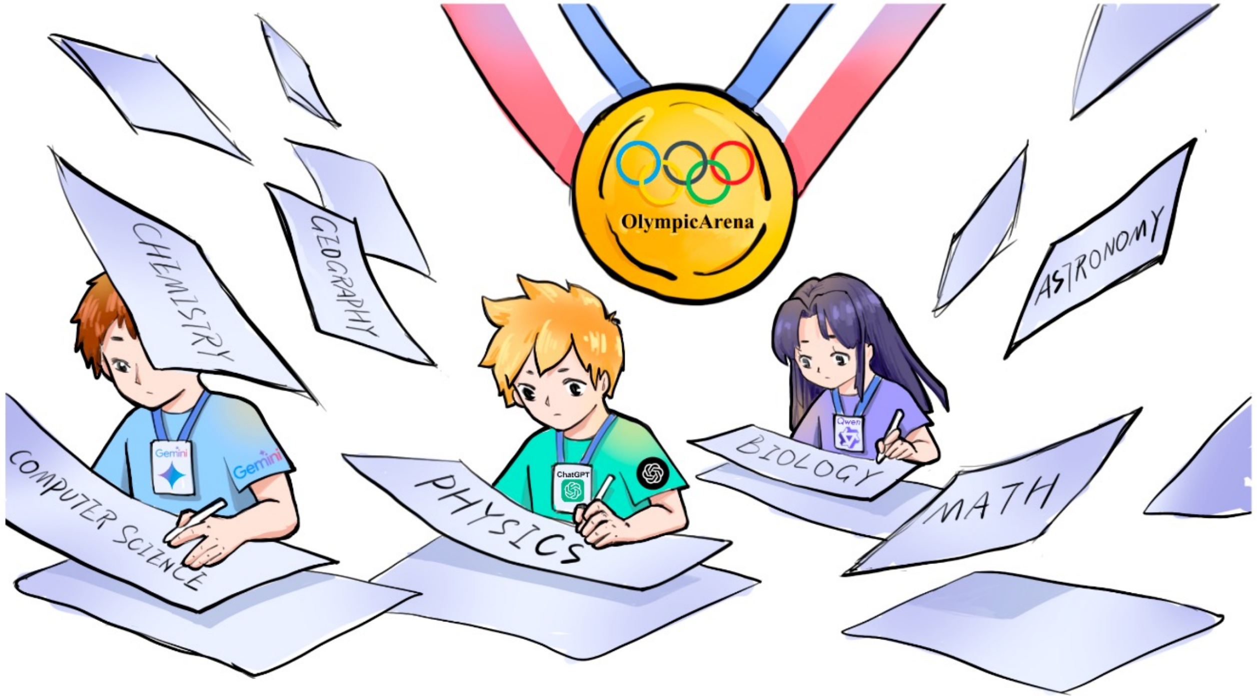
The student's solution

{the model's solution}

Your scores for each step of the student's solutions

OlympicArena





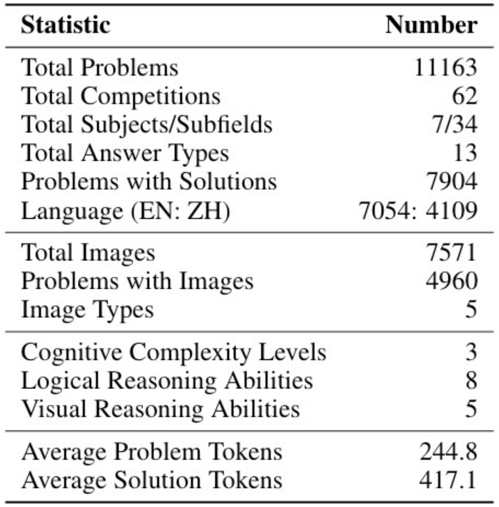
OlympicArena

Data Collection

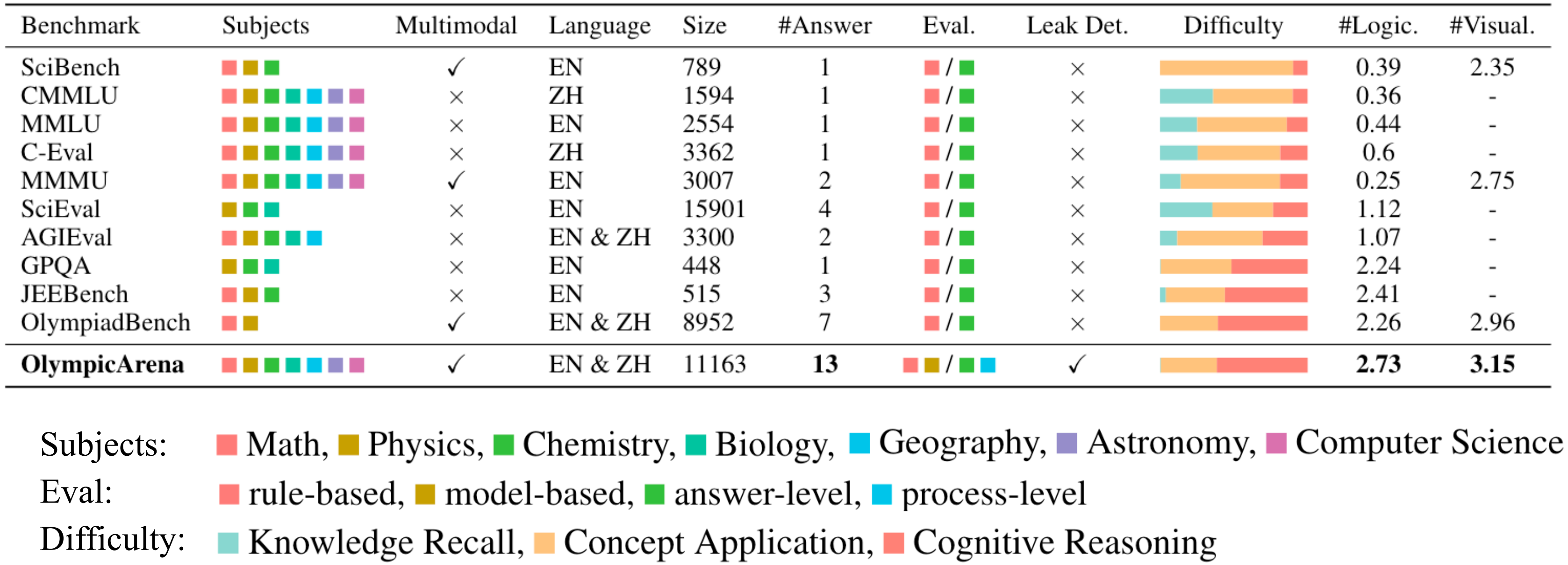
- ✓ Collect URLs of various competitions and download PDFs.
- ✓ Utilize the Mathpix tool to convert PDFs to markdowns.
- ✓ Crawl test cases for CS programming problems.

Data Annotation

- ✓ Develop a user interface and recruit 30 students with STEM background to extract & annotate meta-data.
- ✓ Conduct a multi-step validation process to ensure quality (rule-based & human-based check).
- ✓ Do deduplication within each competition based on model embeddings.
- ✓ Use GPT-4V to annotate difficulty & cognitive reasoning abilities and conduct human verification.



Statistic	Number
Total Problems	11163
Total Competitions	62
Total Subjects/Subfields	7/34
Total Answer Types	13
Problems with Solutions	7904
Language (EN: ZH)	7054: 4109
Total Images	7571
Problems with Images	4960
Image Types	5
Cognitive Complexity Levels	3
Logical Reasoning Abilities	8
Visual Reasoning Abilities	5
Average Problem Tokens	244.8
Average Solution Tokens	417.1



Benchmark	Subjects	Multimodal	Language	Size	#Answer	Eval.	Leak Det.	Difficulty	#Logic.	#Visual.
SciBench	🟥 🟨 🟩	✓	EN	789	1	🟥 / 🟩	×		0.39	2.35
CMMLU	🟥 🟨 🟩 🟢 🟦 🟪 🟫	×	ZH	1594	1	🟥 / 🟩	×		0.36	-
MMLU	🟥 🟨 🟩 🟢 🟦 🟪 🟫	×	EN	2554	1	🟥 / 🟩	×		0.44	-
C-Eval	🟥 🟨 🟩 🟢 🟦 🟪 🟫	×	ZH	3362	1	🟥 / 🟩	×		0.6	-
MMMU	🟥 🟨 🟩 🟢 🟦 🟪 🟫	✓	EN	3007	2	🟥 / 🟩	×		0.25	2.75
SciEval	🟨 🟩 🟢	×	EN	15901	4	🟥 / 🟩	×		1.12	-
AGIEval	🟥 🟨 🟩 🟢 🟦	×	EN & ZH	3300	2	🟥 / 🟩	×		1.07	-
GPQA	🟨 🟩 🟢	×	EN	448	1	🟥 / 🟩	×		2.24	-
JEEBench	🟥 🟨 🟩	×	EN	515	3	🟥 / 🟩	×		2.41	-
OlympiadBench	🟥 🟨	✓	EN & ZH	8952	7	🟥 / 🟩	×		2.26	2.96
OlympicArena	🟥 🟨 🟩 🟢 🟦 🟪 🟫	✓	EN & ZH	11163	**13**	🟥 🟨 / 🟩 🟦	✓		**2.73**	**3.15**

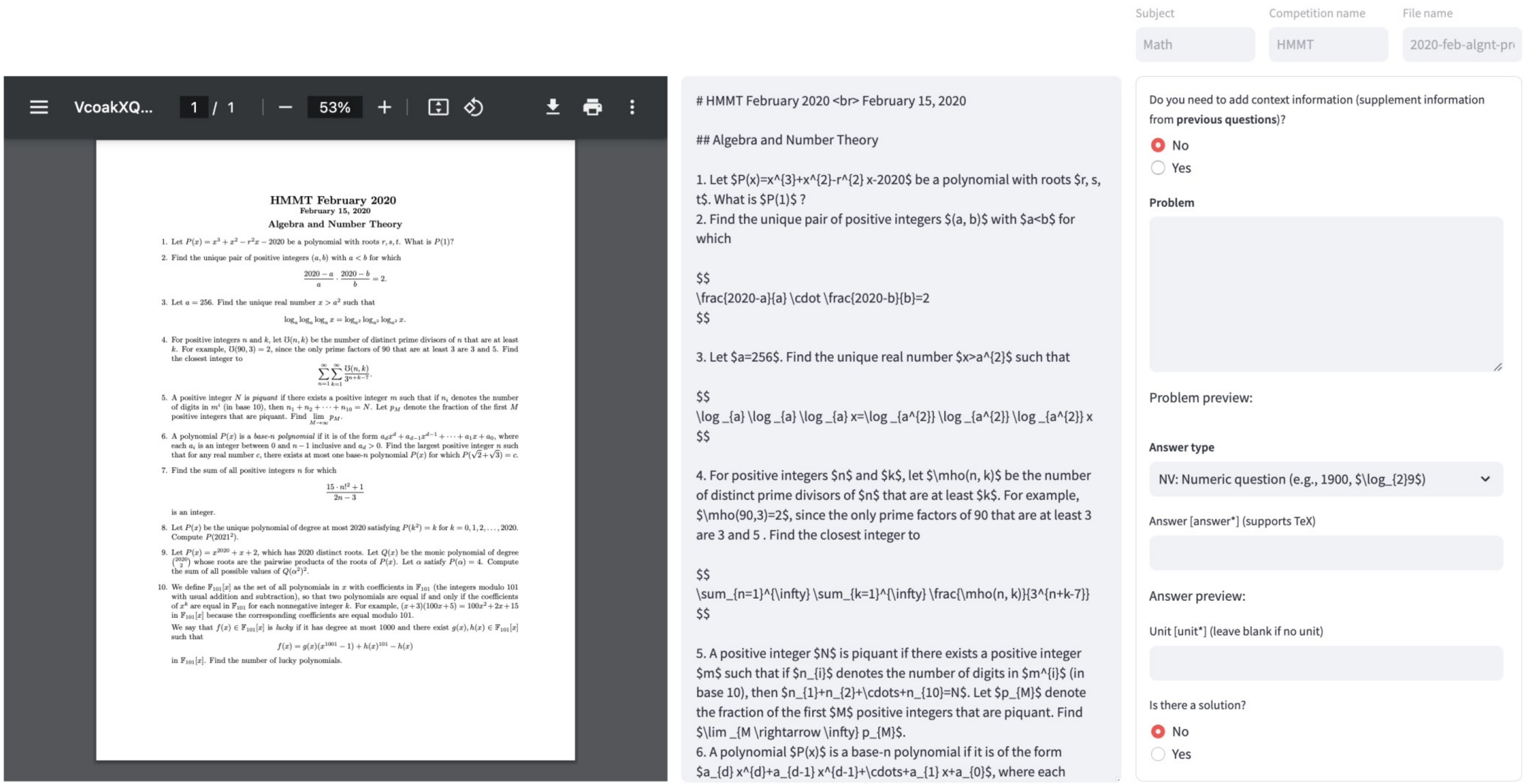
Subjects: 🟥 Math, 🟨 Physics, 🟩 Chemistry, 🟢 Biology, 🟦 Geography, 🟪 Astronomy, 🟫 Computer Science

Eval: 🟥 rule-based, 🟨 model-based, 🟩 answer-level, 🟦 process-level

Difficulty: ■ Knowledge Recall, ■ Concept Application, ■ Cognitive Reasoning

OlympicArena

NEURAL INFORMATION PROCESSING SYSTEMS



HMMT February 2020

February 15, 2020

Algebra and Number Theory

1. Let $P(x)=x^3+x^2-r^2x-2020$ be a polynomial with roots r, s, t. What is $P(1)$?
2. Find the unique pair of positive integers (a,b) with $a<b$ for which
$$\frac{2020-a}{a}\cdot\frac{2020-b}{b}=2.$$
3. Let $a=256$. Find the unique real number $x>a^2$ such that
$$\log_a\log_a\log_a x=\log_{a^2}\log_{a^2}\log_{a^2} x.$$
4. For positive integers n and k, let $\mho(n,k)$ be the number of distinct prime divisors of n that are at least k. For example, $\mho(90,3)=2$, since the only prime factors of 90 that are at least 3 are 3 and 5. Find the closest integer to
$$\sum_{n=1}^{\infty}\sum_{k=1}^{\infty}\frac{\mho(n,k)}{3^{n+k-7}}.$$
5. A positive integer N is *piquant* if there exists a positive integer m such that if n_i denotes the number of digits in m^i (in base 10), then $n_1+n_2+\cdots+n_{10}=N$. Let p_M denote the fraction of the first M positive integers that are piquant. Find $\lim_{M\to\infty}p_M$.
6. A polynomial $P(x)$ is a *base-n polynomial* if it is of the form $a_dx^d+a_{d-1}x^{d-1}+\cdots+a_1x+a_0$, where each a_i is an integer between 0 and $n-1$ inclusive and $a_d>0$. Find the largest positive integer n such that for any real number c, there exists at most one base-n polynomial $P(x)$ for which $P(\sqrt{2}+\sqrt{3})=c$.
7. Find the sum of all positive integers n for which
$$\frac{15\cdot n!^2+1}{2n-3}$$
is an integer.
8. Let $P(x)$ be the unique polynomial of degree at most 2020 satisfying $P(k^2)=k$ for $k=0,1,2,\ldots,2020$. Compute $P(2021^2)$.
9. Let $P(x)=x^{2020}+x+2$, which has 2020 distinct roots. Let $Q(x)$ be the monic polynomial of degree $\binom{2020}{2}$ whose roots are the pairwise products of the roots of $P(x)$. Let α satisfy $P(\alpha)=4$. Compute the sum of all possible values of $Q(\alpha^2)^2$.
10. We define $\mathbb{F}_{101}[x]$ as the set of all polynomials in x with coefficients in $\mathbb{F}_{101}$ (the integers modulo 101 with usual addition and subtraction), so that two polynomials are equal if and only if the coefficients of x^k are equal in $\mathbb{F}_{101}$ for each nonnegative integer k. For example, $(x+3)(100x+5)=100x^2+2x+15$ in $\mathbb{F}_{101}[x]$ because the corresponding coefficients are equal modulo 101.

    We say that $f(x)\in\mathbb{F}_{101}[x]$ is *lucky* if it has degree at most 1000 and there exist $g(x),h(x)\in\mathbb{F}_{101}[x]$ such that
$$f(x)=g(x)(x^{1001}-1)+h(x)^{101}-h(x)$$
in $\mathbb{F}_{101}[x]$. Find the number of lucky polynomials.

```
# HMMT February 2020 <br> February 15, 2020

## Algebra and Number Theory

1. Let $P(x)=x^{3}+x^{2}-r^{2} x-2020$ be a polynomial with roots $r, s, t$. What is $P(1)$ ?
2. Find the unique pair of positive integers $(a, b)$ with $a<b$ for which

$$
\frac{2020-a}{a} \cdot \frac{2020-b}{b}=2
$$

3. Let $a=256$. Find the unique real number $x>a^{2}$ such that

$$
\log _{a} \log _{a} \log _{a} x=\log _{a^{2}} \log _{a^{2}} \log _{a^{2}} x
$$

4. For positive integers $n$ and $k$, let $\mho(n, k)$ be the number of distinct prime divisors of $n$ that are at least $k$. For example, $\mho(90,3)=2$, since the only prime factors of 90 that are at least 3 are 3 and 5 . Find the closest integer to

$$
\sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \frac{\mho(n, k)}{3^{n+k-7}}
$$

5. A positive integer $N$ is piquant if there exists a positive integer $m$ such that if $n_{i}$ denotes the number of digits in $m^{i}$ (in base 10), then $n_{1}+n_{2}+\cdots+n_{10}=N$. Let $p_{M}$ denote the fraction of the first $M$ positive integers that are piquant. Find $\lim _{M \rightarrow \infty} p_{M}$.
6. A polynomial $P(x)$ is a base-n polynomial if it is of the form $a_{d} x^{d}+a_{d-1} x^{d-1}+\cdots+a_{1} x+a_{0}$, where each
```

Subject: Math

Competition name: HMMT

File name: 2020-feb-algnt-pr

Do you need to add context information (supplement information from **previous questions**)?

● No

○ Yes

Problem

Problem preview:

Answer type

NV: Numeric question (e.g., 1900, $\log_{2}9$)

Answer [answer*] (supports TeX)

Answer preview:

Unit [unit*] (leave blank if no unit)

Is there a solution?

● No

○ Yes

Annotation Page

Experiments

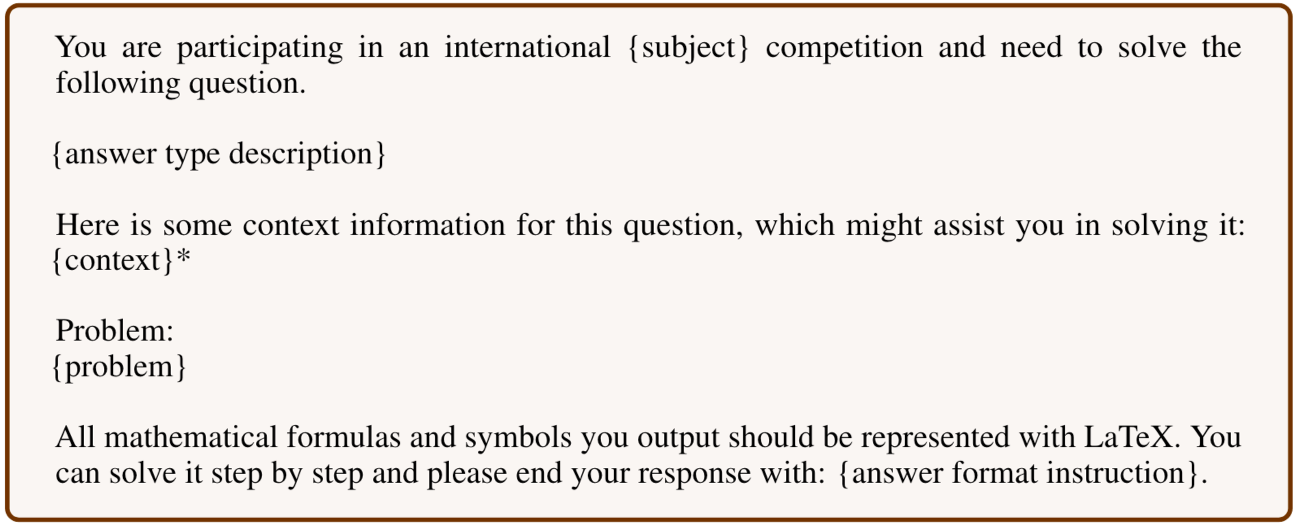
Experimental Setup

□ **Three settings: LLMs, Image caption + LLMs, LMMs**

- **LLMs: w/o any image information**
- **Image caption + LLMs: image -> text description**
- **LMMs: Interleaved image-text input**

Analyze the gains of multimodal information.

□ **Zero-shot CoT prompt (tailored to each answer type)**



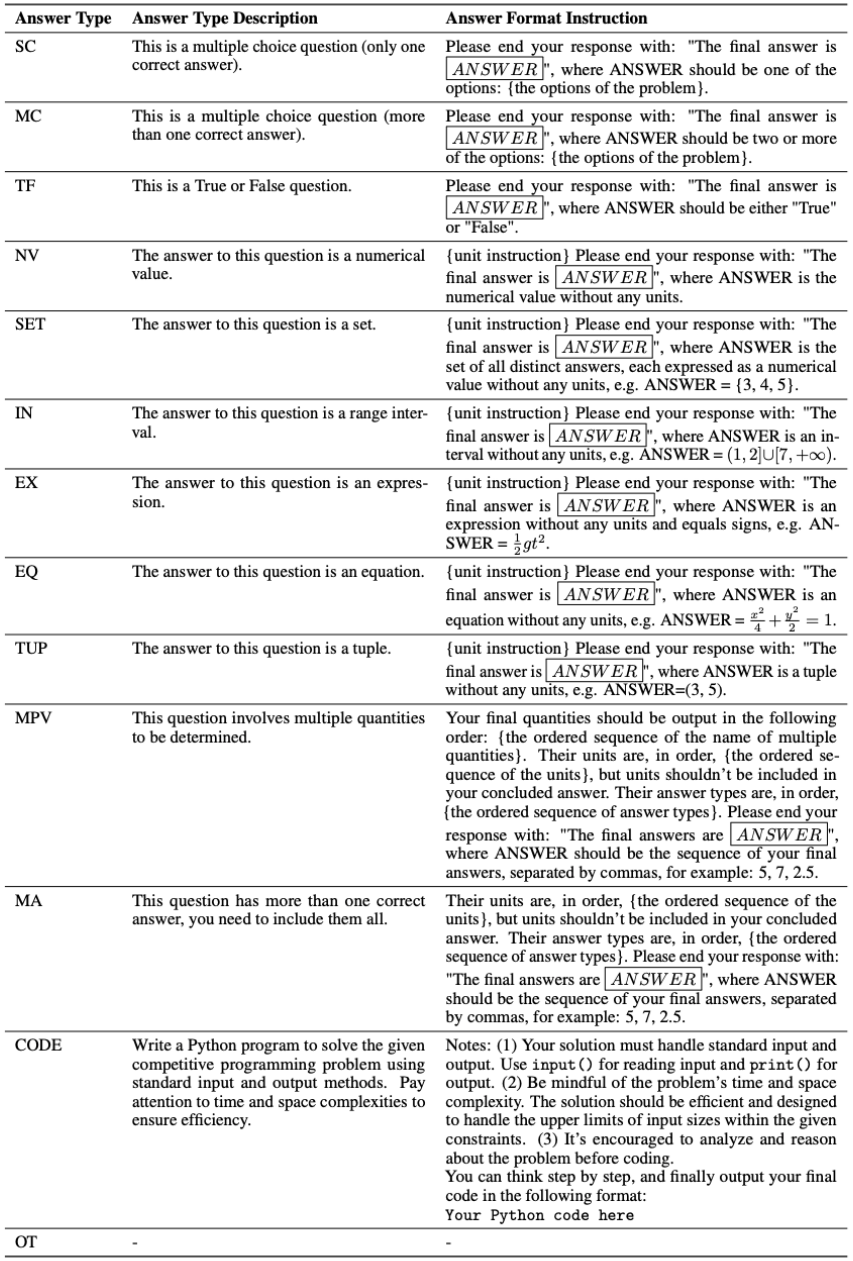
You are participating in an international {subject} competition and need to solve the following question.

{answer type description}

Here is some context information for this question, which might assist you in solving it: {context}*

Problem:
{problem}

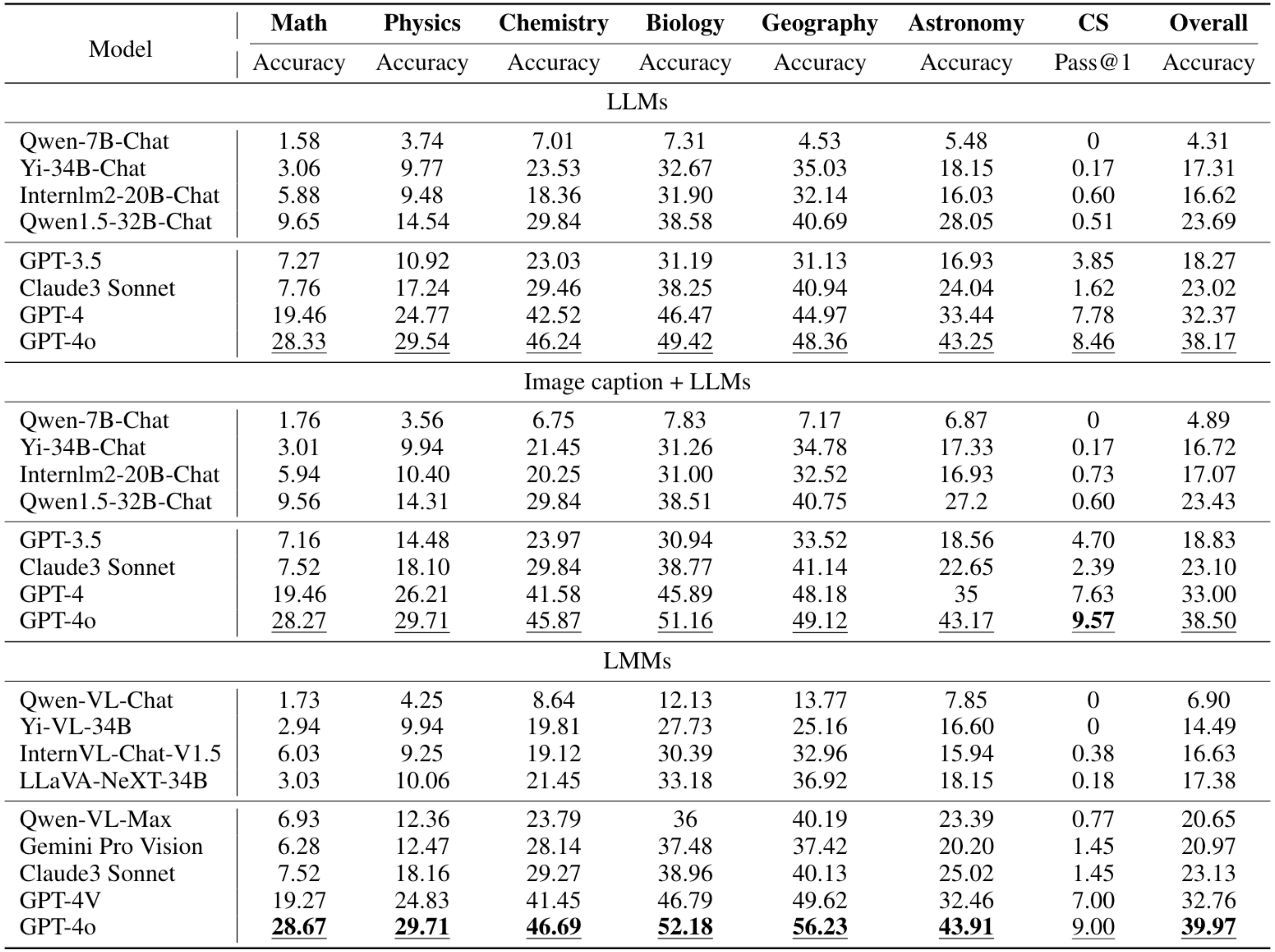
All mathematical formulas and symbols you output should be represented with LaTeX. You can solve it step by step and please end your response with: {answer format instruction}.



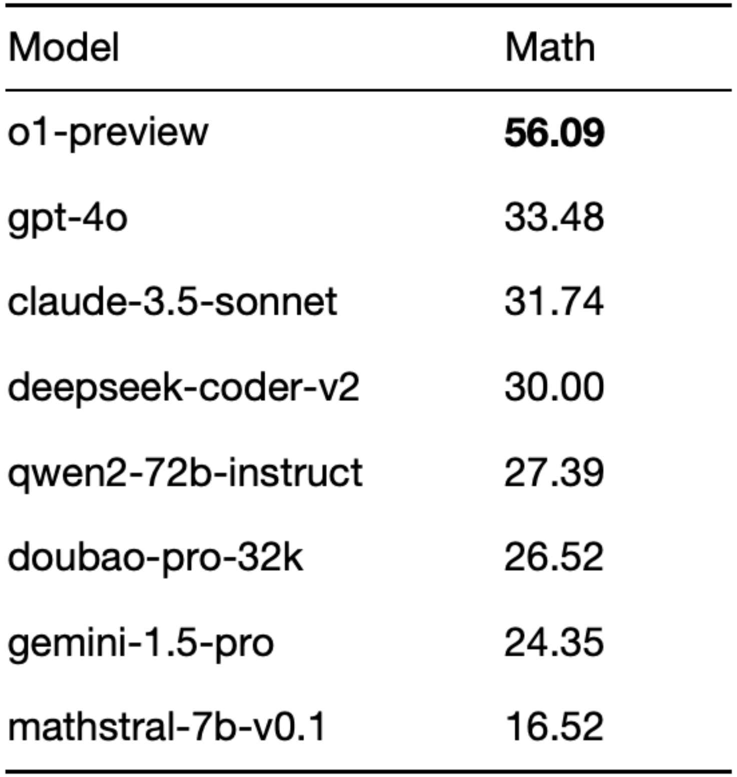
Answer Type	Answer Type Description	Answer Format Instruction
SC	This is a multiple choice question (only one correct answer).	Please end your response with: "The final answer is $\boxed{ANSWER}$", where ANSWER should be one of the options: {the options of the problem}.
MC	This is a multiple choice question (more than one correct answer).	Please end your response with: "The final answer is $\boxed{ANSWER}$", where ANSWER should be two or more of the options: {the options of the problem}.
TF	This is a True or False question.	Please end your response with: "The final answer is $\boxed{ANSWER}$", where ANSWER should be either "True" or "False".
NV	The answer to this question is a numerical value.	{unit instruction} Please end your response with: "The final answer is $\boxed{ANSWER}$", where ANSWER is the numerical value without any units.
SET	The answer to this question is a set.	{unit instruction} Please end your response with: "The final answer is $\boxed{ANSWER}$", where ANSWER is the set of all distinct answers, each expressed as a numerical value without any units, e.g. ANSWER = {3, 4, 5}.
IN	The answer to this question is a range interval.	{unit instruction} Please end your response with: "The final answer is $\boxed{ANSWER}$", where ANSWER is an interval without any units, e.g. ANSWER = $(1, 2] \cup [7, +\infty)$.
EX	The answer to this question is an expression.	{unit instruction} Please end your response with: "The final answer is $\boxed{ANSWER}$", where ANSWER is an expression without any units and equals signs, e.g. ANSWER = $\frac{1}{2}gt^2$.
EQ	The answer to this question is an equation.	{unit instruction} Please end your response with: "The final answer is $\boxed{ANSWER}$", where ANSWER is an equation without any units, e.g. ANSWER = $\frac{x^2}{4} + \frac{y^2}{2} = 1$.
TUP	The answer to this question is a tuple.	{unit instruction} Please end your response with: "The final answer is $\boxed{ANSWER}$", where ANSWER is a tuple without any units, e.g. ANSWER=(3, 5).
MPV	This question involves multiple quantities to be determined.	Your final quantities should be output in the following order: {the ordered sequence of the name of multiple quantities}. Their units are, in order, {the ordered sequence of the units}, but units shouldn't be included in your concluded answer. Their answer types are, in order, {the ordered sequence of answer types}. Please end your response with: "The final answers are $\boxed{ANSWER}$", where ANSWER should be the sequence of your final answers, separated by commas, for example: 5, 7, 2.5.
MA	This question has more than one correct answer, you need to include them all.	Their units are, in order, {the ordered sequence of the units}, but units shouldn't be included in your concluded answer. Their answer types are, in order, {the ordered sequence of answer types}. Please end your response with: "The final answers are $\boxed{ANSWER}$", where ANSWER should be the sequence of your final answers, separated by commas, for example: 5, 7, 2.5.
CODE	Write a Python program to solve the given competitive programming problem using standard input and output methods. Pay attention to time and space complexities to ensure efficiency.	Notes: (1) Your solution must handle standard input and output. Use `input()` for reading input and `print()` for output. (2) Be mindful of the problem's time and space complexity. The solution should be efficient and designed to handle the upper limits of input sizes within the given constraints. (3) It's encouraged to analyze and reason about the problem before coding. You can think step by step, and finally output your final code in the following format: `Your Python code here`
OT	-	-

Experiments

Main Results



Model	Math	Physics	Chemistry	Biology	Geography	Astronomy	CS	Overall
	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Pass@1	Accuracy
LLMs								
Qwen-7B-Chat	1.58	3.74	7.01	7.31	4.53	5.48	0	4.31
Yi-34B-Chat	3.06	9.77	23.53	32.67	35.03	18.15	0.17	17.31
Internlm2-20B-Chat	5.88	9.48	18.36	31.90	32.14	16.03	0.60	16.62
Qwen1.5-32B-Chat	9.65	14.54	29.84	38.58	40.69	28.05	0.51	23.69
GPT-3.5	7.27	10.92	23.03	31.19	31.13	16.93	3.85	18.27
Claude3 Sonnet	7.76	17.24	29.46	38.25	40.94	24.04	1.62	23.02
GPT-4	19.46	24.77	42.52	46.47	44.97	33.44	7.78	32.37
GPT-4o	28.33	29.54	46.24	49.42	48.36	43.25	8.46	38.17
Image caption + LLMs								
Qwen-7B-Chat	1.76	3.56	6.75	7.83	7.17	6.87	0	4.89
Yi-34B-Chat	3.01	9.94	21.45	31.26	34.78	17.33	0.17	16.72
Internlm2-20B-Chat	5.94	10.40	20.25	31.00	32.52	16.93	0.73	17.07
Qwen1.5-32B-Chat	9.56	14.31	29.84	38.51	40.75	27.2	0.60	23.43
GPT-3.5	7.16	14.48	23.97	30.94	33.52	18.56	4.70	18.83
Claude3 Sonnet	7.52	18.10	29.84	38.77	41.14	22.65	2.39	23.10
GPT-4	19.46	26.21	41.58	45.89	48.18	35	7.63	33.00
GPT-4o	28.27	29.71	45.87	51.16	49.12	43.17	**9.57**	38.50
LMMs								
Qwen-VL-Chat	1.73	4.25	8.64	12.13	13.77	7.85	0	6.90
Yi-VL-34B	2.94	9.94	19.81	27.73	25.16	16.60	0	14.49
InternVL-Chat-V1.5	6.03	9.25	19.12	30.39	32.96	15.94	0.38	16.63
LLaVA-NeXT-34B	3.03	10.06	21.45	33.18	36.92	18.15	0.18	17.38
Qwen-VL-Max	6.93	12.36	23.79	36	40.19	23.39	0.77	20.65
Gemini Pro Vision	6.28	12.47	28.14	37.48	37.42	20.20	1.45	20.97
Claude3 Sonnet	7.52	18.16	29.27	38.96	40.13	25.02	1.45	23.13
GPT-4V	19.27	24.83	41.45	46.79	49.62	32.46	7.00	32.76
GPT-4o	**28.67**	**29.71**	**46.69**	**52.18**	**56.23**	**43.91**	9.00	**39.97**



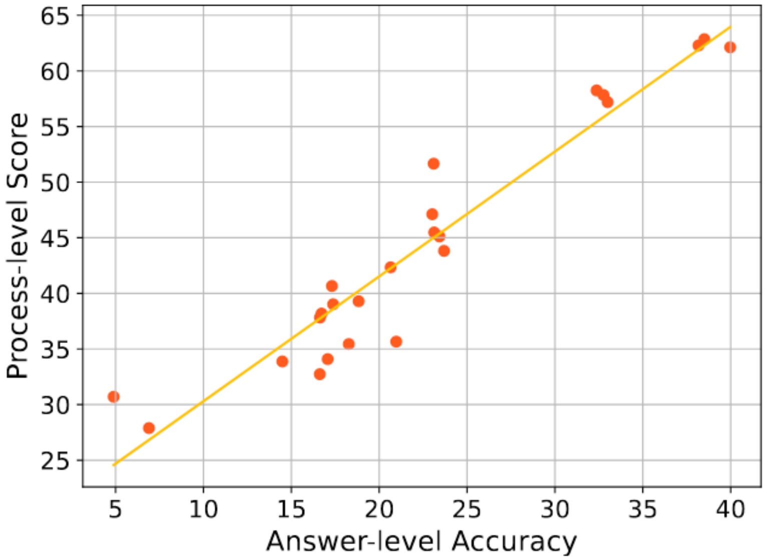
Model	Math
o1-preview	**56.09**
gpt-4o	33.48
claude-3.5-sonnet	31.74
deepseek-coder-v2	30.00
qwen2-72b-instruct	27.39
doubao-pro-32k	26.52
gemini-1.5-pro	24.35
mathstral-7b-v0.1	16.52

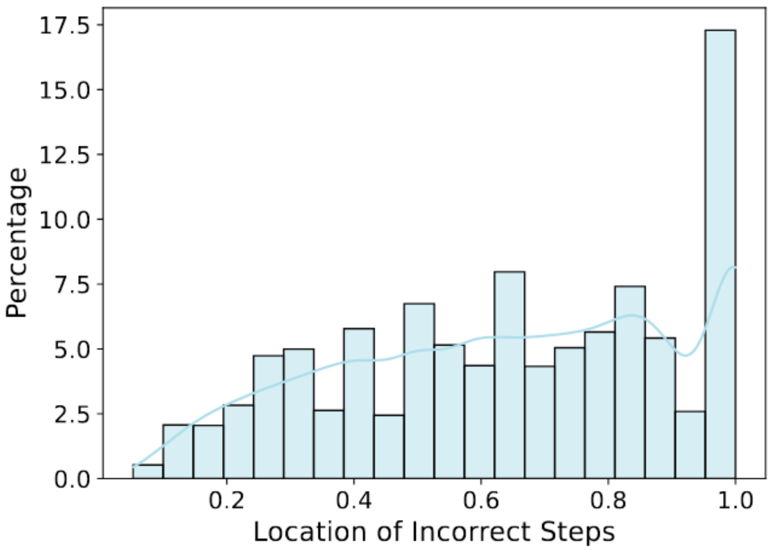
OlympicArena Math Problems validation set (text-only)

Experiments

Fine-grained Analysis

- Analysis of process-level evaluation results





□ There is generally a high consistency between process-level evaluation and answer-level evaluation.

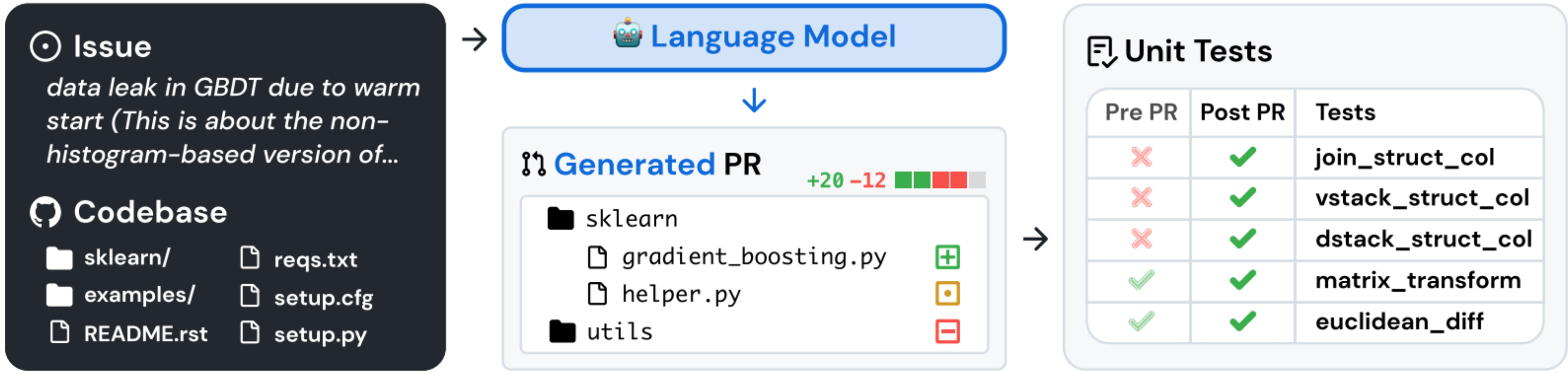
□ The accuracy at the process-level is often higher than at the answer-level.

□ A higher proportion of errors occur in the later stages.

Discussion

- Is using Olympiads to benchmark AI sufficient?

From problem-solving to tackling **real-world tasks (AI4Science, AI4SE, etc.)**



Issue

data leak in GBDT due to warm start (This is about the non-histogram-based version of...

Codebase

sklearn/ reqs.txt
examples/ setup.cfg
README.rst setup.py

Language Model

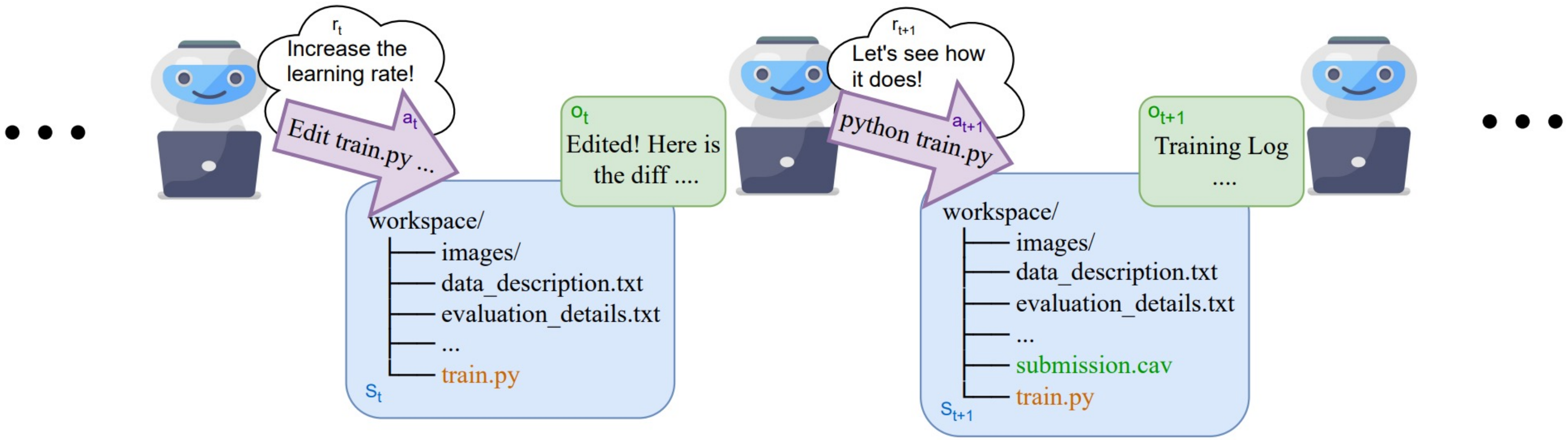
Generated PR +20 −12

```
sklearn
    gradient_boosting.py
    helper.py
utils
```

Unit Tests

Pre PR	Post PR	Tests
✗	✓	join_struct_col
✗	✓	vstack_struct_col
✗	✓	dstack_struct_col
✓	✓	matrix_transform
✓	✓	euclidean_diff

SWE-Bench



```
workspace/
├── images/
├── data_description.txt
├── evaluation_details.txt
├── ...
└── train.py
```
s_t

```
workspace/
├── images/
├── data_description.txt
├── evaluation_details.txt
├── ...
├── submission.cav
└── train.py
```
s_{t+1}

MLAgentBench

Thanks.