https://mathvista.github.io

MathVista: Evaluating Math Reasoning in Visual Contexts with GPT-4V, Bard, and Other Large Multimodal Models

Pan Lu[1], Hritik Bansal[1], Tony Xia[1], Jiacheng Liu[2], Chunyuan Li[3],
Hannaneh Hajishirzi[2], Hao Cheng[3], Kai-Wei Chang[1], Michel Galley[3], Jianfeng Gao[3]

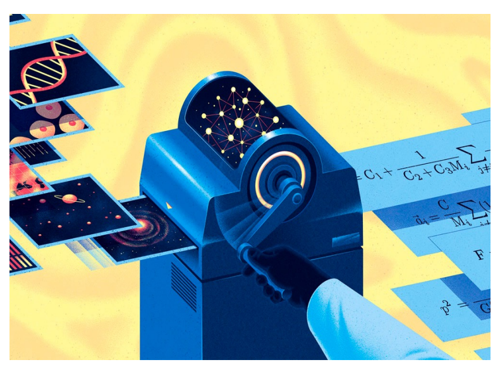
[1]UCLA [2]University of Washington [3]Microsoft Research

MathVista: Why Math Reasoning in Visual Contexts?

Mesopotamia, 2000BC

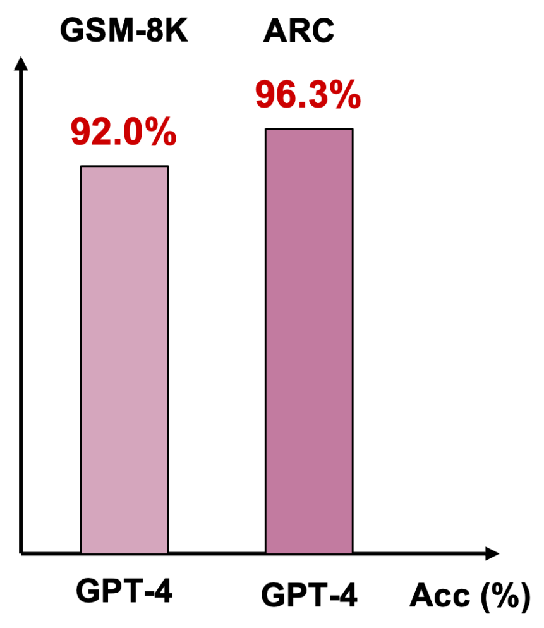
K-12 Math Education

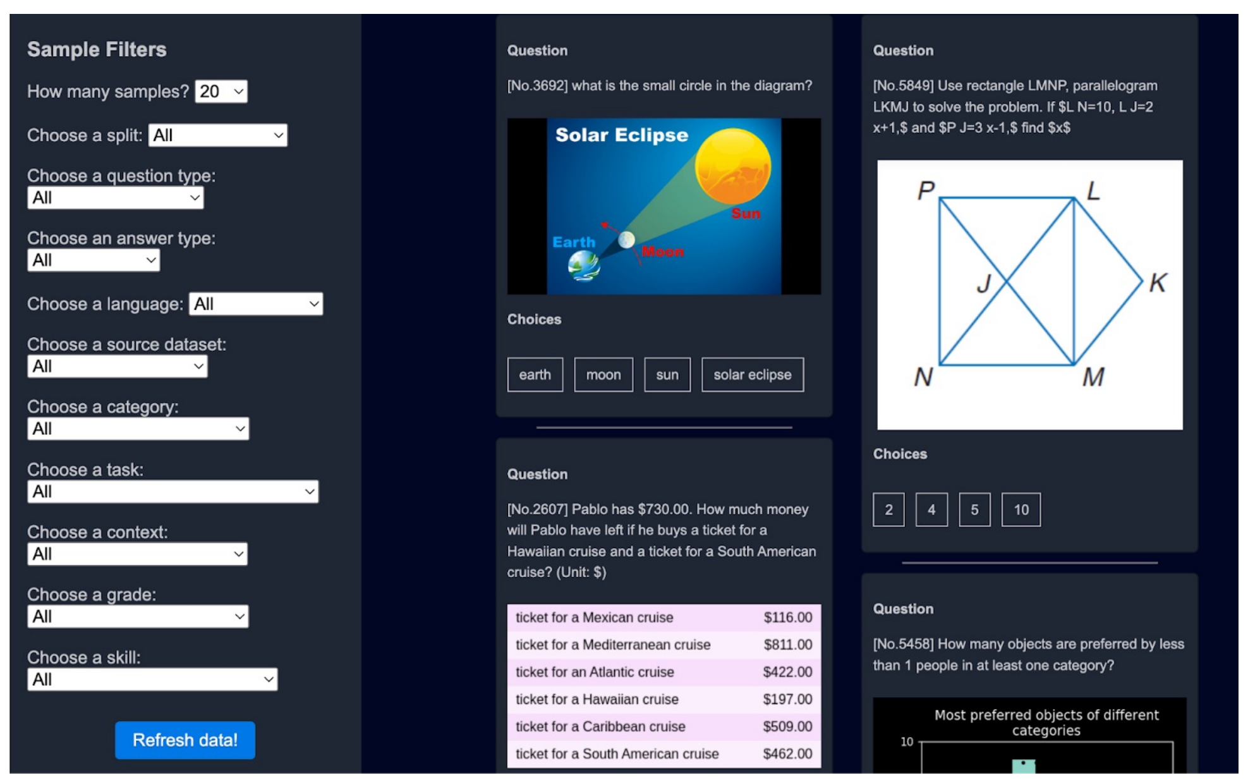
Data Analysis


Scientific Discovery

Existing Challenges

- Text-only Benchmarks
- Narrow-scope Benchmarks
- Benchmark Saturation
- Missing Systematic Evaluations





Question

[No.3692] what is the small circle in the diagram?

Choices

earth | moon | sun | solar eclipse

Question

[No.2607] Pablo has $730.00. How much money will Pablo have left if he buys a ticket for a Hawaiian cruise and a ticket for a South American cruise? (Unit: $)

ticket for a Mexican cruise	$116.00
ticket for a Mediterranean cruise	$811.00
ticket for an Atlantic cruise	$422.00
ticket for a Hawaiian cruise	$197.00
ticket for a Caribbean cruise	$509.00
ticket for a South American cruise	$462.00

Question

[No.5849] Use rectangle LMNP, parallelogram LKMJ to solve the problem. If $L N=10, L J=2 x+1,$ and $P J=3 x-1,$ find x

Choices

2 | 4 | 5 | 10

Question

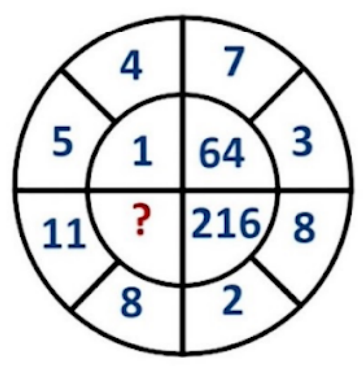
[No.5458] How many objects are preferred by less than 1 people in at least one category?

The creation of a novel testbed is **essential** to

- **Facilitate** the development of **mathematical reasoning systems** in visually intensive scenarios
- **Evaluate** the research progress of **foundation models**, especially regarding their capabilities in rigorous reasoning

The MathVista Benchmark

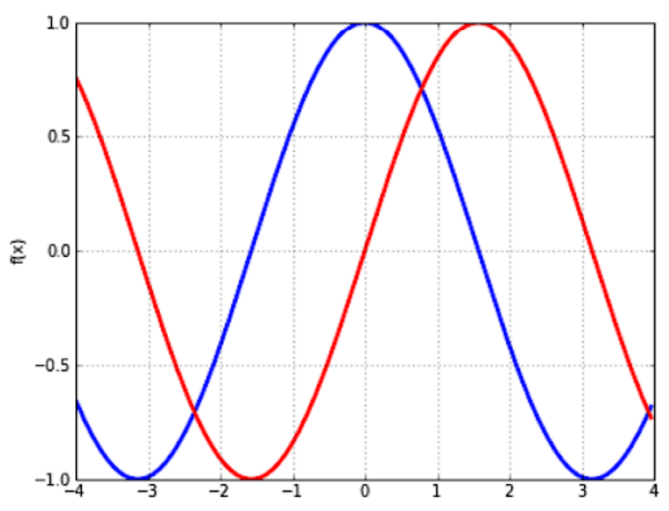
Statistic	Number
Total questions	6,141
- multiple-choice questions	3,392 (55.2%)
- Free-form questions	2,749 (44.8%)
- Questions with annotations	5,261 (85.6%)
- Questions newly annotated	736 (12.0%)
Unique number of images	5,487
Unique number of questions	4,746
Unique number of answers	1,464
Source datasets	31
- Existing VQA datasets	19
- Existing MathQA datasets	9
- Our newly annotated datasets	3
Visual context (image) classes	19
Maximum question length	213
Maximum answer length	27
Maximum choice number	8
Average question length	15.6
Average answer length	1.2
Average choice number	3.4



Question: Find the missing value in this math puzzle.
Solution:
$(5 - 4)^3 = 1$
$(7 - 3)^3 = 64$
$(8 - 2)^3 = 216$
Similarly, $(11 - 8)^3 = 27$.
So the missing value is 27.
Answer: 27

Category: Math-targeted
Task: Figure question answering
Context: Puzzle test
Grade: Elementary school
Math: Logical reasoning

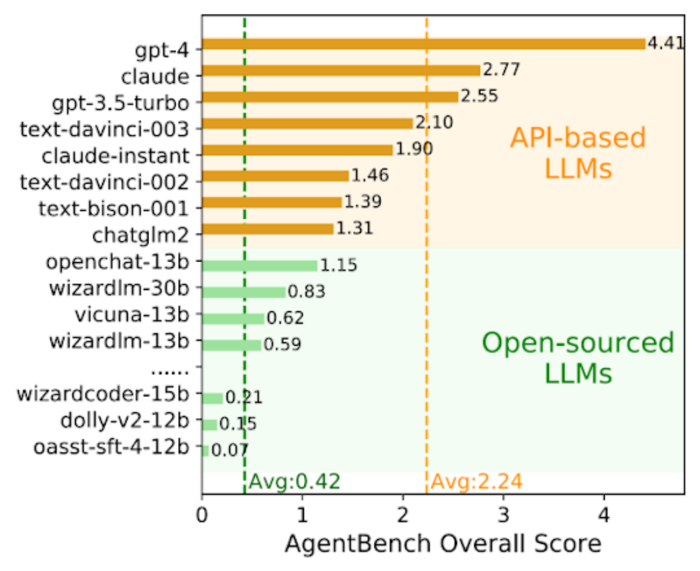
(a) IQTest



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

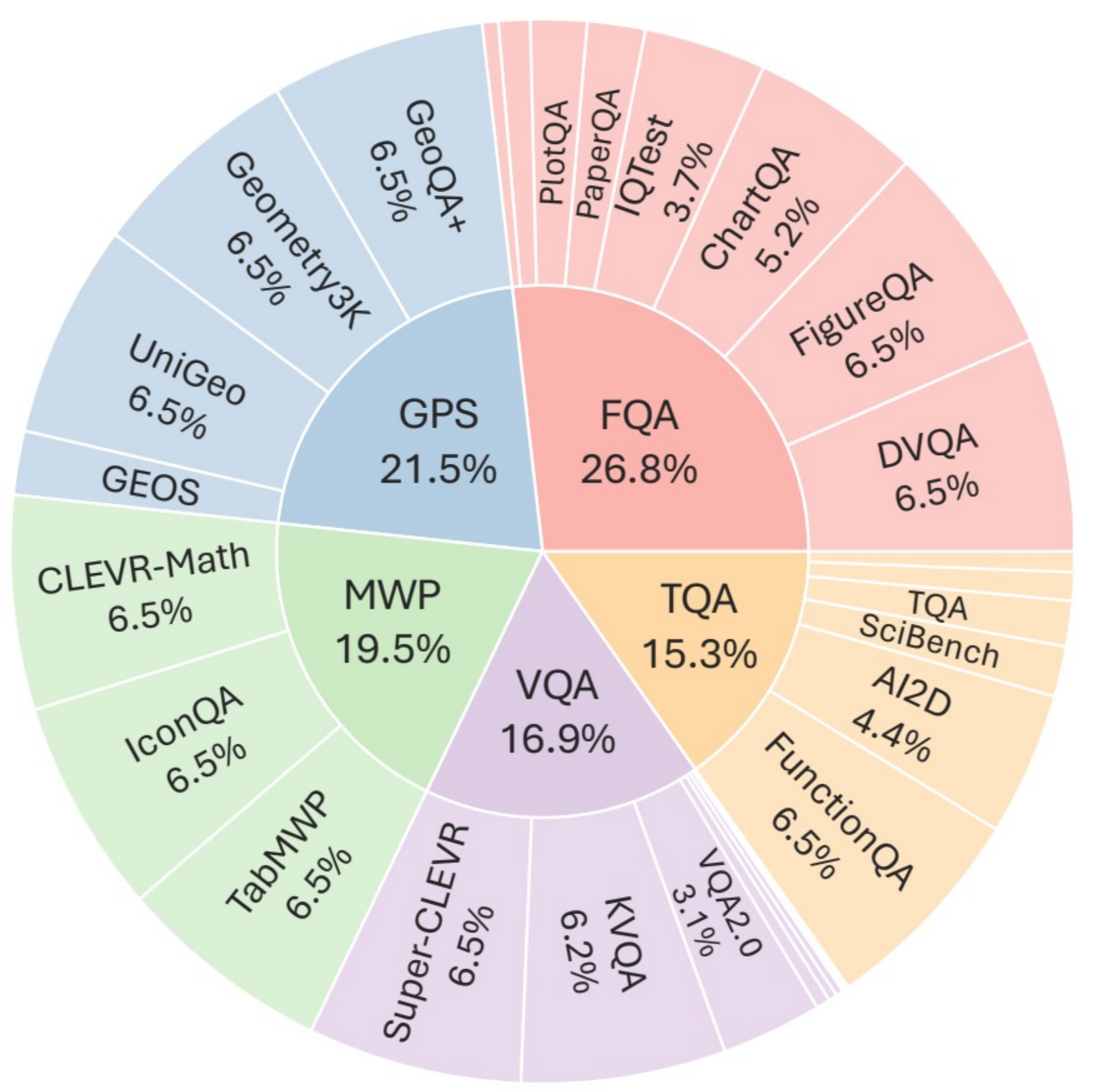
(b) FunctionQA



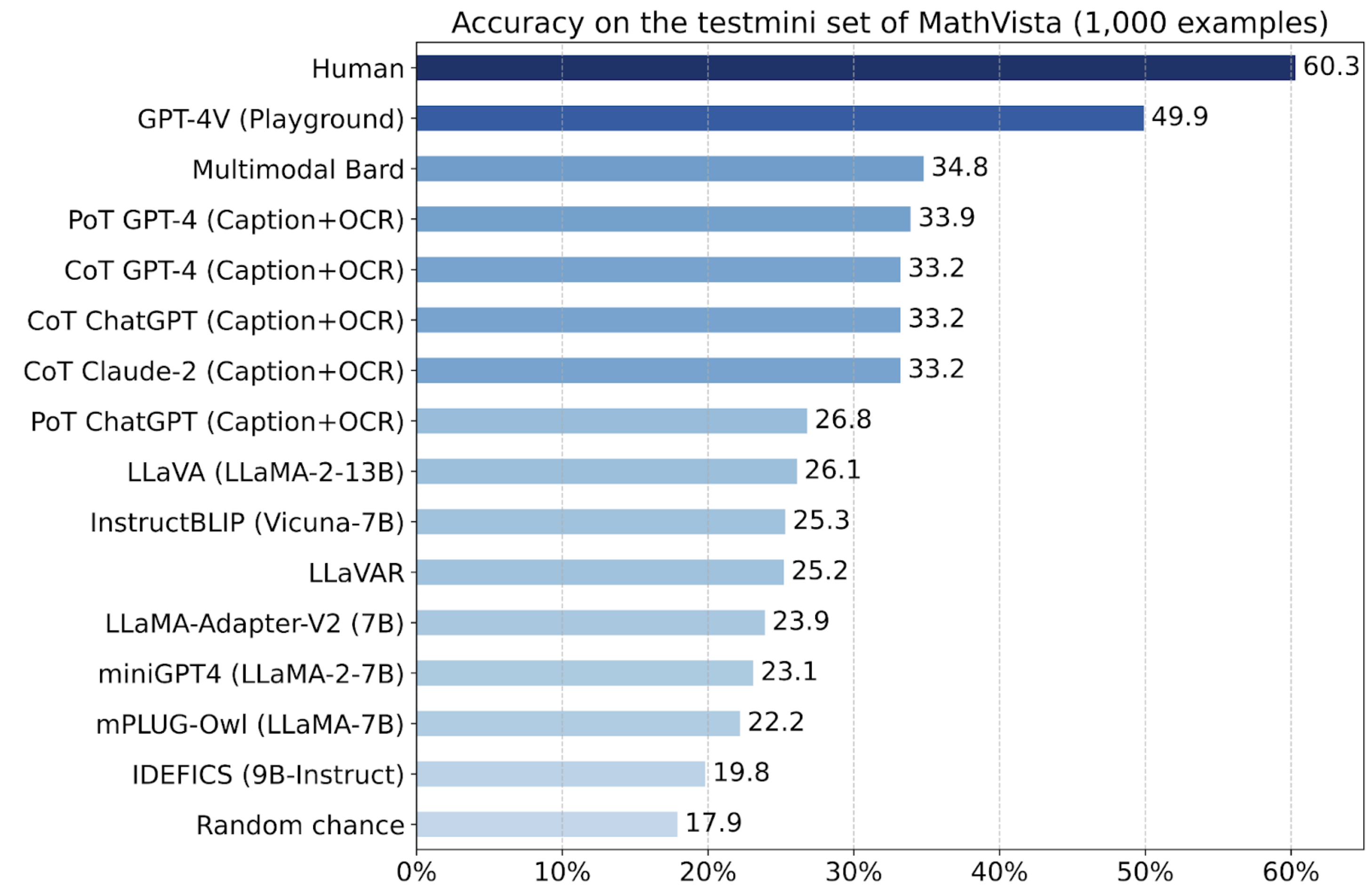
Question: What is the performance gap in the AgentBench Overall Score between the worst API-based LLM and the best open-sourced LLM?
Answer: 0.16

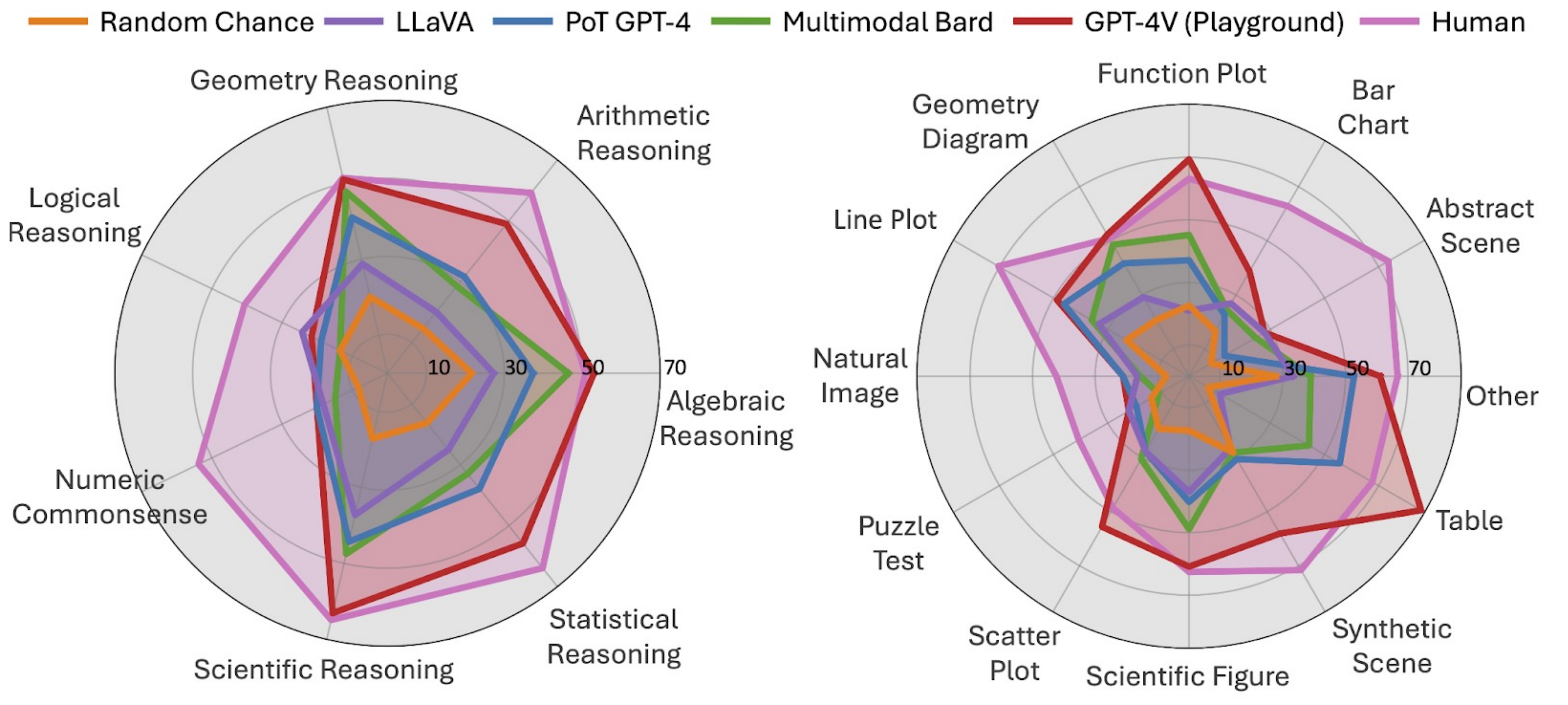
Category: Math-targeted
Task: Figure question answering
Context: Scientific figure
Grade: College
Math: Scientific reasoning

(c) PaperQA



Quantitative Evaluations with 12 Foundation Models (GPT-4V, Bard, and More)





(a) Mathematical reasoning (b) Visual context

Augmented LLM (GPT-4)

Laps driven	
Day	Number of laps
Thursday	53
Friday	53
Saturday	53
Sunday	55
Monday	55

Question: A race car driver kept track of how many laps he drove in the past 5 days. What is the mode of the numbers?

Detected OCR text:
([55, 9], Laps driven),
([21, 34], Day),
([81, 37], Number of laps),
([1, 64], Thursday), ([129, 67], 53),
([3, 99], Friday), ([129, 99], 53),
([1, 126], Saturday), ([129, 129], 53),
([3, 161], Sunday), ([129, 161], 55),
([1, 188], Monday), ([129, 191], 55)

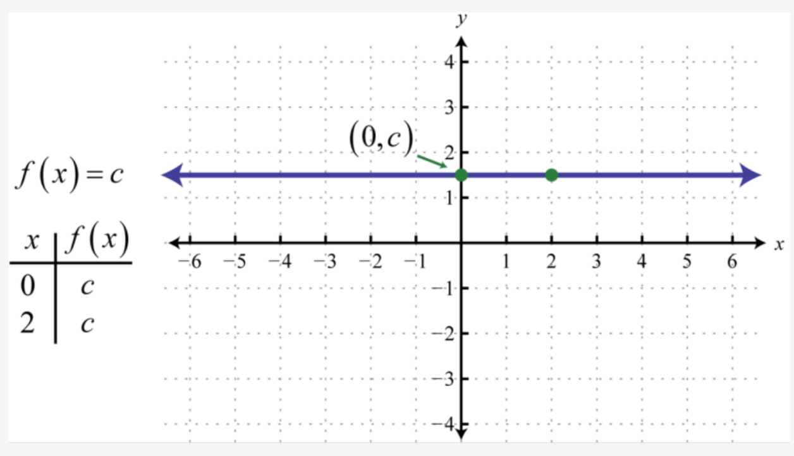
Response from GPT-4:

```
def mode(laps):
    return max(set(laps), key
   =laps.count)

laps = [53, 53, 53, 55, 55]
print(mode(laps))
```

Correct output: 53

(a) **Correct answer** and **code**



Question: What is this function most likely be?
Choices:
(A) polynomial (B) trigonometric
(C) exponential (D) logarithmic

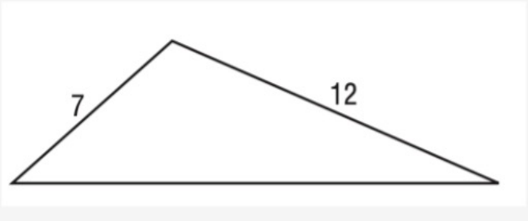
Caption from Bard:
The image shows a graph of a function $f(x) = c$ on a graph paper. The x axis is the horizontal axis and the y axis is the vertical axis. The graph is a horizontal line that passes through the point $(0, c)$. ... The text "$y = 3$" means that the y-value of all points on the graph is 3. The text "$(0, c)$" means that the point $(0, c)$ is on the graph. The text "$f(x) = c$" means that the function $f(x)$ is equal to c for all values of x. ... The graph of a constant function is a horizontal line. In this case, the horizontal line passes through the point $(0, 3)$, so the constant value of the function is 3. ...

Response from GPT-4:
The function shown in the image is a constant function, which is a type of polynomial function. Thus, the answer is A.

Correct output: (A) polynomial

(b) **Correct answer** with **partially correct outputs**

Self-Verification in GPT-4V



Question: If two sides of a triangle measure 12 and 7, which of the following cannot be the perimeter of the triangle?

Choices: (A) 29 (B) 34 (C) 37 (D) 38

Ground truth: (D) 38

GPT-4V:
The sum of the lengths of any two sides of a triangle must be greater than the length of the third side. Let x be the length of the third side.

From the triangle inequality:

$$12 + 7 > x => x < 19$$

Similarly:

$$7 + x > 12 => x > 5$$

And:

$$12 + x > 7 => x > -5$$

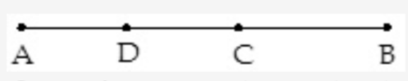
(This will always be true since the length cannot be negative)

So,

$$5 < x < 19$$

Now, let's check the given options:

(A) 29 => $x = 29 - 12 - 7 = 10$ (Possible)
(B) 34 => $x = 34 - 12 - 7 = 15$ (Possible)
(C) 37 => $x = 37 - 12 - 7 = 18$ (Possible)
(D) 38 => $x = 38 - 12 - 7 = 19$ (Not Possible)
Correct output: (D) 38



Question: As shown in the figure, if $CB = 4.0$, $DB = 7.0$, and D is the midpoint of AC, then the length of AC is ().
Choices: (A) 6 cm (B) 7 cm (C) 4 cm (D) 5 cm
Ground truth: (A) 6 cm

GPT-4V:

...

$$AC/2 = -3.0$$

This is not possible as lengths cannot be negative. Let's re-evaluate our approach.

From the original figure:

$$AB = AD + DB$$

$$AB = DC + DB$$

Given $DB = 7.0$ cm, we have:

$$AD = AB - 7.0$$

$$AD = AC/2 - 7.0$$

Using the relation $AB = AC + CB$:

$$AC/2 + 7.0 = AC + 4.0$$

Subtracting $AC/2$ from both sides:

$$7.0 = AC/2 + 4.0$$

Subtracting 4 from both sides:

$$AC/2 = 3.0$$

Multiplying both sides by 2:

$$AC = 6.0 \text{ cm}$$

Correct output: (A) 6 cm

Self-Consistency for GPT-4V

Web pages visited	
Person	Number of Web pages
Bonnie	84
Winston	56
Percy	70
Derek	68
Jake	60
Finn	47
Carter	75
Kurt	52

ID: 331

Question: Several people compared how many Web pages they had visited. What is the mean of the numbers?

Ground truth: 64

GPT-4V (Reasoning Path 1):
To find the mean of the numbers, you add up all the numbers and then divide by the quantity of numbers.

$$\begin{aligned}\text{Mean} &= (84 + 56 + 70 + 68 + 60 + 47 + 75 + 52) \div 8\\ &= 462 \div 8\\ &= 57.75\end{aligned}$$

However, since the question requires an integer answer, the mean number of web pages visited is: 58.
Wrong output: 58

GPT-4V (Reasoning Path 2):
To find the mean number of web pages visited:

First, sum the number of web pages for all the individuals:
Bonnie: 84, Winston: 56, Percy: 70, Derek: 68, Jake: 60, Finn: 47, Carter: 75, Kurt: 52

$$\text{Total} = 84 + 56 + 70 + 68 + 60 + 47 + 75 + 52 = 512$$

Next, divide by the number of individuals (8): Mean $= 512 \div 8 = 64$

The mean number of web pages visited is 64.
Correct output: 64

GPT-4V (Reasoning Path 3):
First, I'll list the number of web pages each person visited:
Bonnie: 84, Winston: 56, Percy: 70, Derek: 68, Jake: 60, Finn: 47, Carter: 75, Kurt: 52

Next, I'll group some numbers together for easier addition:

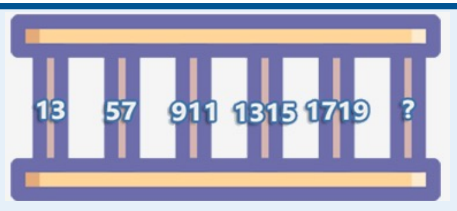
$$(84 + 56) = 140$$
$$(70 + 68) = 138$$
$$(60 + 47) = 107$$
$$(75 + 52) = 127$$

Now, I'll sum these results: $140 + 138 + 107 + 127 = 512$

Now, I'll divide the sum by the number of individuals to get the mean: $512 \div 8 = 64$

So, the mean number of web pages visited is 64.
Correct output: 64

Multi-turn Human-AI Interaction in GPT-4V



Which number comes next?

The pattern seems to involve prime numbers. Let's break it down: 1 and 3 are the first and second prime numbers. 5 and 7 are the third and fourth prime numbers. 9 (not a prime) but followed by 11 which is the fifth prime number. 13 is the sixth prime number and 15 (not a prime) followed by 17 which is the seventh prime number. 19 is the eighth prime number.

Following this pattern, the next two prime numbers are 23 and 29. So, the next number in the sequence would be 2329.

Your solution is incorrect. The pattern does not involve prime numbers.

I apologize for the oversight. Let's re-examine the pattern. It seems the pattern is derived by concatenating consecutive numbers: 1 and 3 give 13. 5 and 7 give 57. 9 and 11 give 911. 13 and 15 give 1315. 17 and 19 give 1719.

Following this pattern, the next numbers to be concatenated would be 21 and 23. So, the next number in the sequence would be 2123.