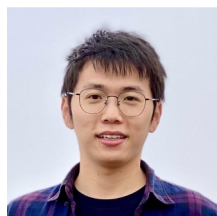




UNIVERSITY of
WASHINGTON



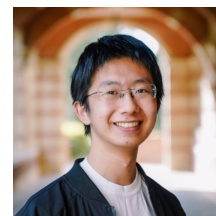
MathVista: Evaluating **Mathematical** Reasoning of Foundation Models in **Visual** Contexts



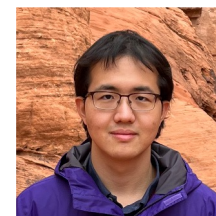
Pan Lu



Hritik Bansal



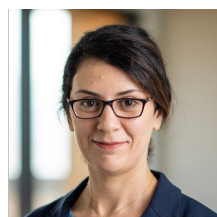
Tony Xia



Jiacheng Liu



Chunyuan Li



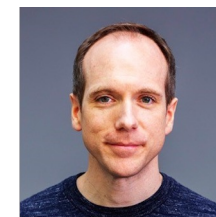
Hannaneh Hajishirzi



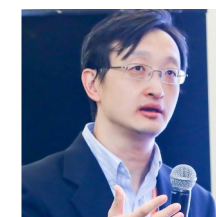
Hao Cheng



Kai-Wei Chang

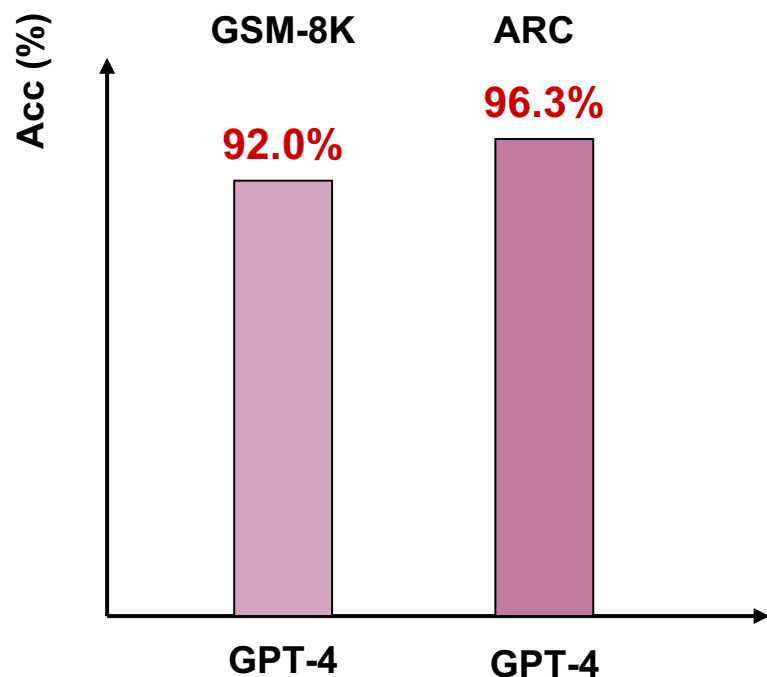


Michel Galley



Jianfeng Gao

Latest Models Potentially Strong Enough



GSM-8K: Grade-school mathematics questions

ARC: Grade-school MC science questions

Problem

The battery charge in Mary's cordless vacuum cleaner lasts ten minutes. It takes her four minutes to vacuum each room in her house. Mary has three bedrooms, a kitchen, and a living room. How many times does Mary need to charge her vacuum cleaner to vacuum her whole house?

Solution

Mary has $3 + 1 + 1 = 5$ rooms in her house.
At 4 minutes a room, it will take her $4 * 5 = 20$ minutes to vacuum her whole house.
At 10 minutes a charge, she will need to charge her vacuum cleaner $20 / 10 = 2$ times to vacuum her whole house.

Final Answer

2

GSM-8K Example

Problem

George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?

Choice

- (A) dry palms
- (B) wet palms
- (C) palms covered with oil
- (D) palms covered with lotion

Answer

(A) dry palms

ARC Example

However, the Road Ahead Remains Long



| Tour boat schedule | | | | | |
|--------------------|------------|------------|------------|------------|------------|
| Ocean City | 8:15 A.M. | 9:00 A.M. | 9:15 A.M. | 9:30 A.M. | 10:00 A.M. |
| Whale Watch Harbor | 9:30 A.M. | 10:15 A.M. | 10:30 A.M. | 10:45 A.M. | 11:15 A.M. |
| Oyster Lighthouse | 10:15 A.M. | 11:00 A.M. | 11:15 A.M. | 11:30 A.M. | 12:00 P.M. |
| Fisherman's Cove | 11:15 A.M. | 12:00 P.M. | 12:15 P.M. | 12:30 P.M. | 1:00 P.M. |
| Surfing Beach | 12:00 P.M. | 12:45 P.M. | 1:00 P.M. | 1:15 P.M. | 1:45 P.M. |

Question: Look at the following schedule. Haley is at Ocean City at 9:45 A.M. How soon can she get to Surfing Beach?

Options: (A) 1:45 P.M. (B) 12:00 P.M. (C) 10:30 A.M. (D) 12:15 P.M.

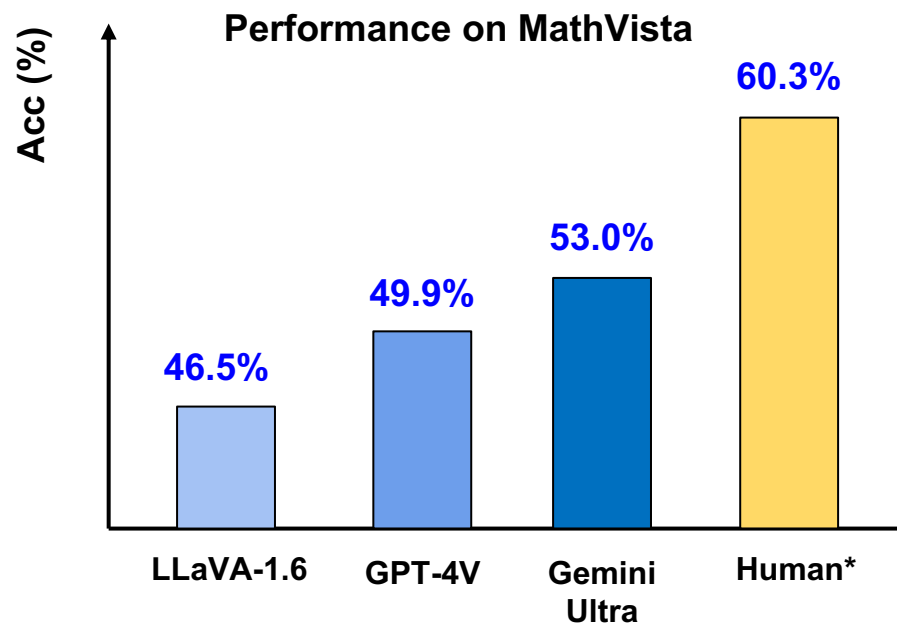


From the schedule, if Haley is at Ocean City at 9:45 A.M., the next boat from Ocean City is at 10:00 A.M. Following the schedule for Surfing Beach, the boat from Ocean City at 10:00 A.M. will reach Surfing Beach at 12:15 P.M.

So, the answer is:
(D) 12:15 P.M.

Hallucination

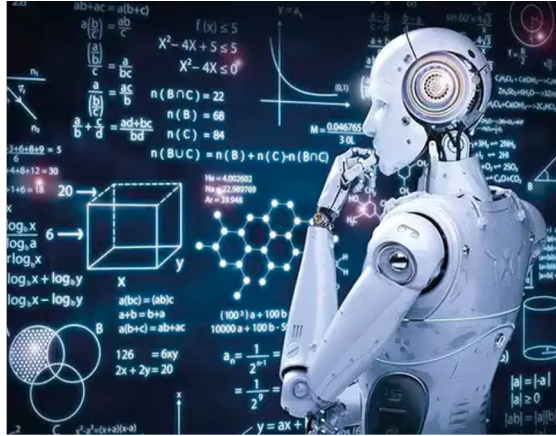
GPT-4V fails on this simple problem



MathVista (2024): math reasoning in visual context (difficulty: from K-12 to college level)

Human*: annotators with high-school diploma or higher

Visual Math Reasoning



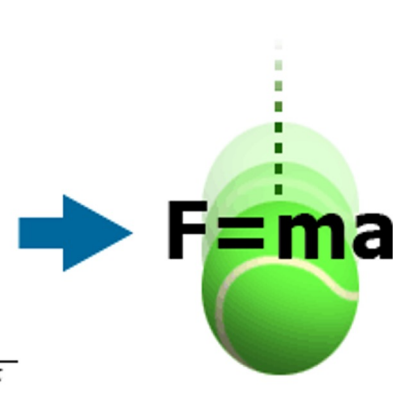
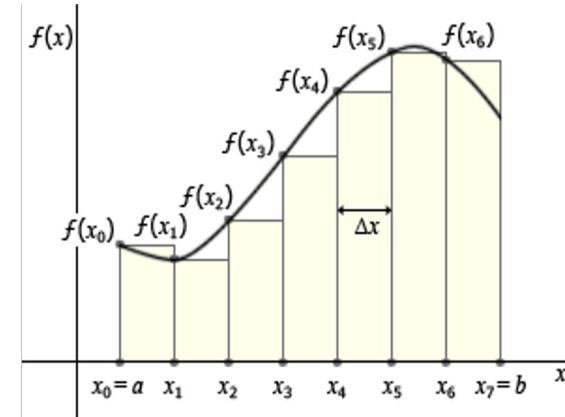
Fundamental AI Problem



2000BC, Mesopotamia



K-12 Math Education



Wide Applications



With MathVista, we report, for the first time, a comprehensive **quantitative** and **qualitative** evaluation of GPT-4V and other large models in **mathematical reasoning in visual contexts**



The MathVista Benchmark

Demo: <https://mathvista.github.io/#visualization>

| 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 |

MathVista Visualizer

Sample Filters

How many samples?

Choose a split:

Choose a question type:

Choose an answer type:

Choose a language:

Choose a source dataset:

Choose a category:

Choose a task:

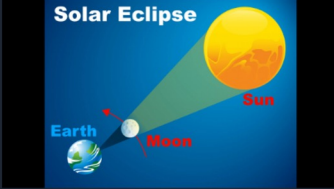
Choose a context:

Choose a grade:

Choose a skill:

Question

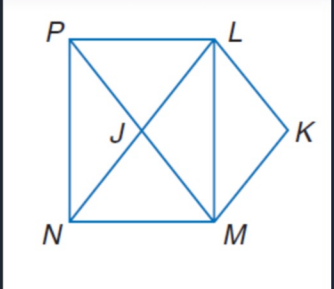
[No.3692] what is the small circle in the diagram?



Choices

Question

[No.5849] Use rectangle LMNP, parallelogram LKMJ to solve the problem. If $SL = 10$, $LJ = 2x + 1$, and $PK = 3x - 1$, find Sx



Choices

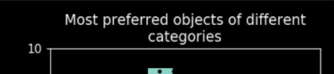
Question

[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

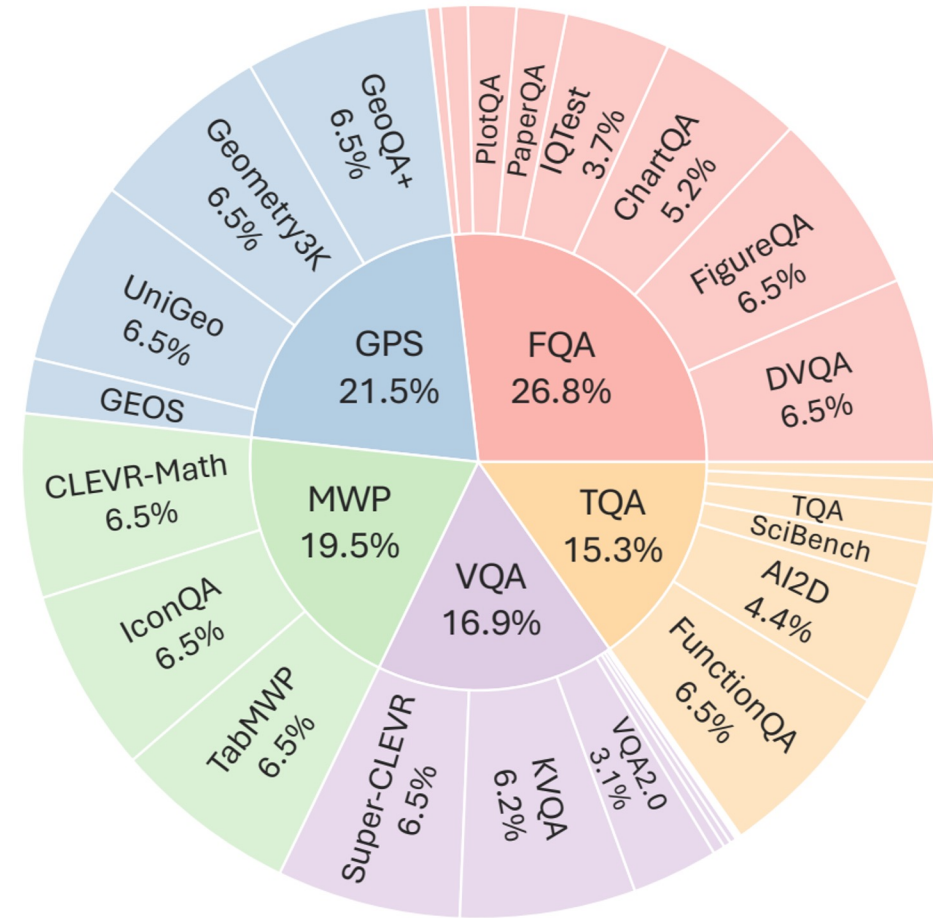
[No.5458] How many objects are preferred by less than 1 people in at least one category?





Comprehensive Sources: 31 Datasets

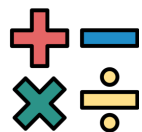
| Dataset | Category | Task | Context | Math Skill |
|---------------------|---------------|------|-------------------|-------------------------|
| IQTest (Ours) | Math-Targeted | FQA | Puzzle Test | Logical, Arithmetic |
| PaperQA (Ours) | Math-Targeted | FQA | Charts and Plots | Scientific |
| FunctionQA (Ours) | Math-Targeted | TQA | Function Plot | Algebraic |
| Geometry3K (2021a) | Math-Targeted | GPS | Geometry Diagram | Geometry, Algebraic |
| GeoQA+ (2022) | Math-Targeted | GPS | Geometry Diagram | Geometry, Algebraic |
| GEOS (2015) | Math-Targeted | GPS | Geometry Diagram | Geometry, Algebraic |
| UniGeo (2022a) | Math-Targeted | GPS | Geometry Diagram | Geometry, Algebraic |
| CLEVR-Math (2022) | Math-Targeted | MWP | Synthetic Scene | Arithmetic |
| IconQA (2021b) | Math-Targeted | MWP | Abstract Scene | Arithmetic |
| TabMWP (2023b) | Math-Targeted | MWP | Table | Statistical, Arithmetic |
| SciBench (2023) | Math-Targeted | TQA | Scientific Figure | Scientific |
| TheoremQA (2023) | Math-Targeted | TQA | Scientific Figure | Scientific |
| ChartQA (2022) | General VQA | FQA | Charts and Plots | Statistical |
| FigureQA (2017) | General VQA | FQA | Charts and Plots | Statistical |
| DVQA (2018) | General VQA | FQA | Bar Chart | Statistical |
| MapQA (2022) | General VQA | FQA | Map Chart | Statistical |
| PlotQA (2020) | General VQA | FQA | Scatter Plot | Statistical |
| DocVQA (2022) | General VQA | FQA | Document Image | Statistical |
| AI2D (2016) | General VQA | TQA | Scientific Figure | Scientific |
| ScienceQA (2022) | General VQA | TQA | Scientific Figure | Scientific |
| TQA (2017) | General VQA | TQA | Scientific Figure | Scientific |
| A-OKVQA (2022) | General VQA | VQA | Natural Image | Arithmetic, Numeric |
| KVQA (2019) | General VQA | VQA | Natural Image | Arithmetic, Numeric |
| ParsVQA-Caps (2022) | General VQA | VQA | Natural Image | Arithmetic, Numeric |
| TextVQA (2019) | General VQA | VQA | Natural Image | Arithmetic, Numeric |
| VizWiz (2018) | General VQA | VQA | Natural Image | Arithmetic, Numeric |
| VQA2.0 (2017) | General VQA | VQA | Natural Image | Arithmetic, Numeric |
| PMC-VQA (2023b) | General VQA | VQA | Medical Image | Scientific |
| VQA-RAD (2018) | General VQA | VQA | Medical Image | Scientific |
| Super-CLEVR (2023b) | General VQA | VQA | Synthetic Scene | Arithmetic |
| VQA-AS (2015) | General VQA | VQA | Abstract Scene | Arithmetic |



Core Math Reasoning



Arithmetic



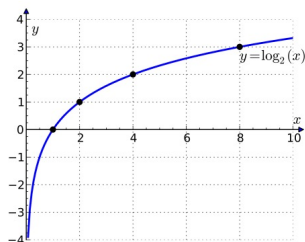
| | |
|-----------------|------------|
| silk scraps | \$9.08/lb |
| denim scraps | \$8.47/lb |
| canvas scraps | \$8.17/lb |
| felt scraps | \$7.29/lb |
| faux fur scraps | \$11.79/lb |
| lace scraps | \$6.37/lb |

Question: Karen bought 4 pounds of silk scraps and 4 pounds of canvas scraps. How much did she spend? (Unit: \$)

Solution:

Find the cost of the silk scraps. Multiply: $\$9.08 \times 4 = \36.32
 Find the cost of the canvas scraps. Multiply: $\$8.17 \times 4 = \32.68
 Now find the total cost by adding: $\$36.32 + \$32.68 = \$69$
 She spent \$69.
Answer: 69

Algebraic



Question: The derivative of y at $x = 6$ is ___ that at $x = 8$.

Choices: (A) larger than (B) equal to (C) smaller than
Answer: (A) larger than

Question: How many zeros does this function have?
Answer: 1

Question: What is the value of y at $x = 1$?
Answer: 0

Numeric



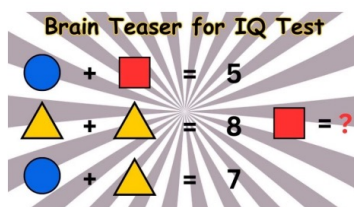
Question: What is the age gap between these two people in image? (unit: years)

Named entities: Winston Churchill, Charles de Gaulle

Wiki caption: Winston Churchill and General de Gaulle at Marrakesh, January 1944

Answer: 16

Logical

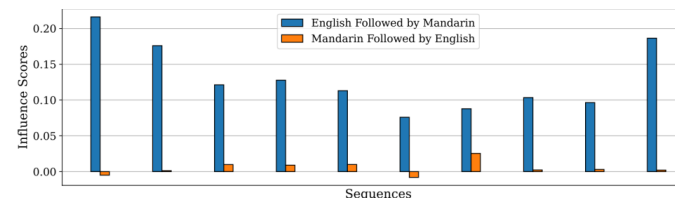


Question: Find the value of the square in the figure.

Solution:

Circle + Square = 5, Triangle + Triangle = 8,
 Triangle = 4.
 Circle + Triangle = 7, Circle = 3.
 Therefore Square = 2

Answer: 2

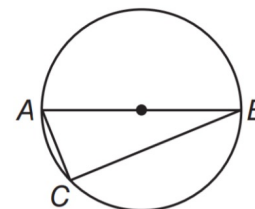


Question: How many sequences have negative Influence Scores?
Answer: 2

Statistical



Geometry

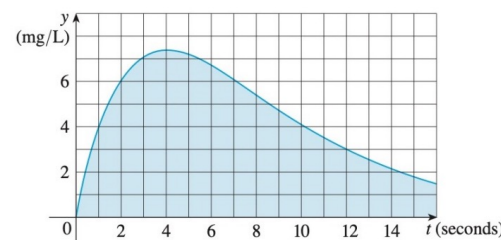


Question: \overline{AB} is a diameter, $AC = 8$ inches, and $BC = 15$ inches. Find the radius of the circle.

Diagram logic forms:

PointLiesOnLine(D, Line(B, A))
 PointLiesOnCircle(B, Circle(D, radius))
 PointLiesOnCircle(A, Circle(D, radius))
 PointLiesOnCircle(C, Circle(D, radius))

Answer: (C) 8.5



Question: The graph of the concentration function $c(t)$ is shown after a 7-mg injection of dye into a heart. Use Simpson's Rule to estimate the cardiac output.

Answer: 5.77

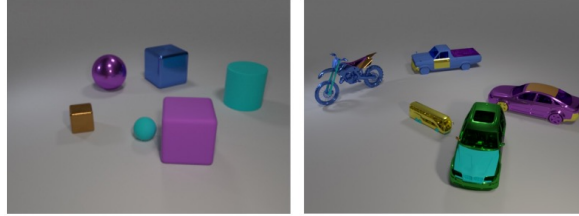
Scientific



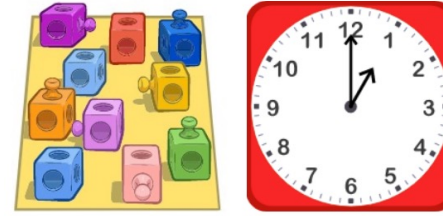
Diverse Visual Contexts



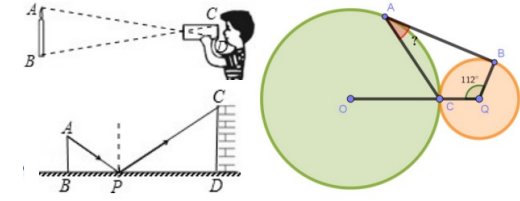
Natural Images



Synthetic Scene



Abstract Scene

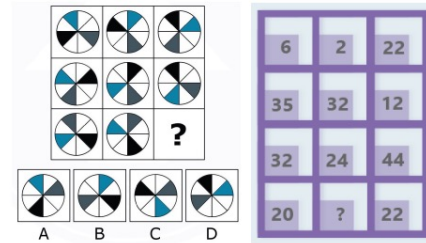


Geometry Diagram

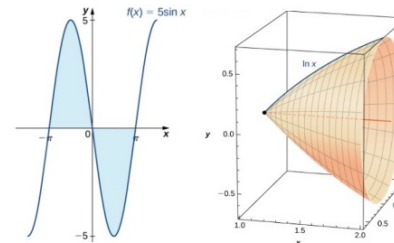
| Cans of food collected | |
|------------------------|------------------------|
| Name | Number of cans of food |
| Emmett | 8 |
| Luther | 7 |
| Bruce | 10 |
| Scott | 9 |
| Mabel | 9 |
| Roxanne | 5 |
| Kevin | 8 |

| Planet | Semimajor Axis a (10^{10} m) | Period T (y) | T^2/a^3 (10^{-34} y^2/m^3) |
|---------|-----------------------------------|----------------|------------------------------------|
| Mercury | 5.79 | 0.241 | 2.99 |
| Venus | 10.8 | 0.615 | 3.00 |
| Earth | 15.0 | 1.00 | 2.96 |
| Mars | 22.8 | 1.88 | 2.98 |
| Jupiter | 77.8 | 11.9 | 3.01 |
| Saturn | 143 | 29.5 | 2.98 |
| Uranus | 287 | 84.0 | 2.98 |
| Neptune | 450 | 165 | 2.99 |
| Pluto | 590 | 248 | 2.99 |

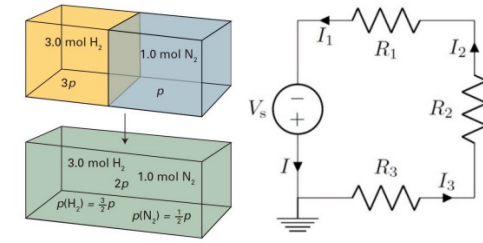
Table



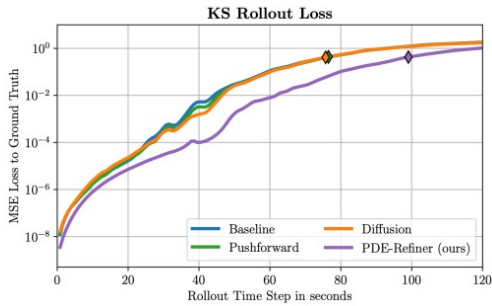
Puzzle Test



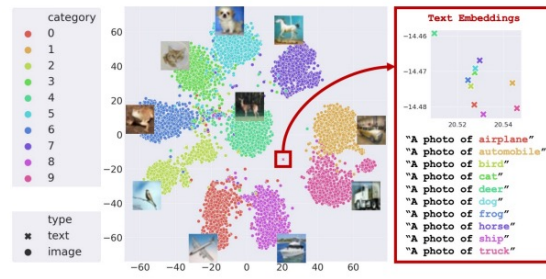
Function Plot



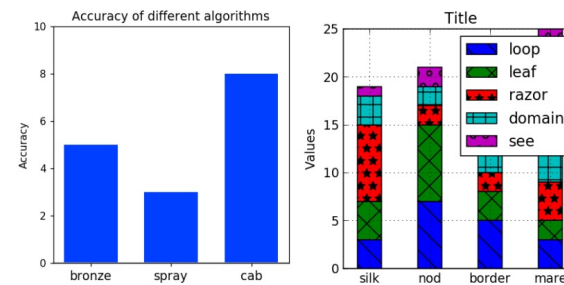
Scientific Figure



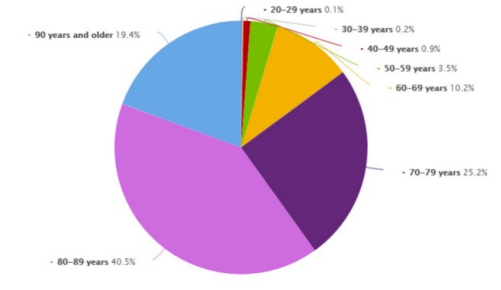
Line Plot



Bar Chart

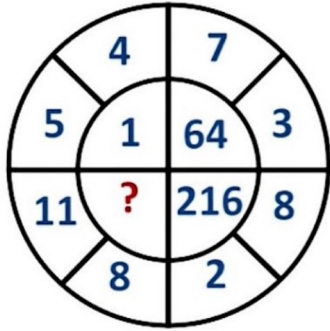


Scatter Plot



Pie Chart

We Created Three New Datasets



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

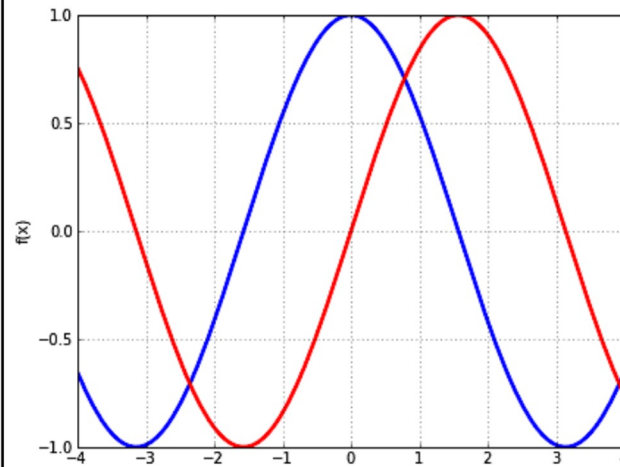
Task: Figure question answering

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

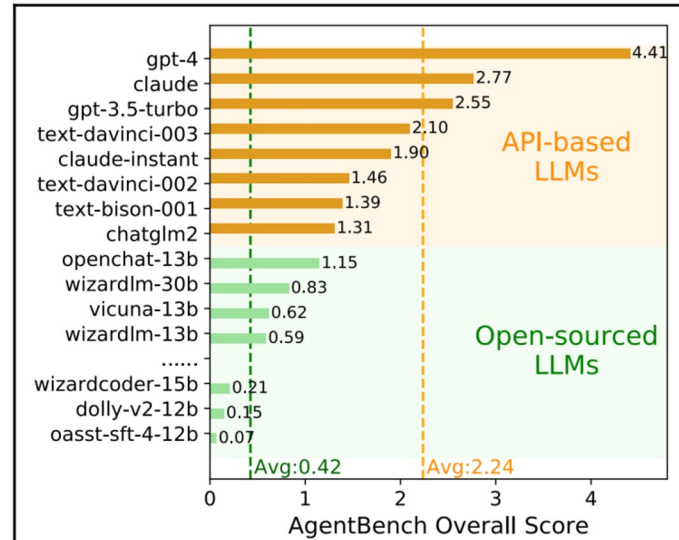
Task: Textbook question answering

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

How Good are Current Advanced Models?



Question: What is the highest amount this glass measures?

Ground truth: 400 ml

Google Bard:

The highest amount the beaker in the image measures is 600 ml.

The markings on the side of the beaker are in milliliters (ml). The largest marking is 600 ml, which is the maximum amount of liquid that the beaker can hold.

The other markings on the beaker are 300 ml, 200 ml, 100 ml, and -400 ml. The negative markings indicate the amount of liquid that is below the zero mark.



Google Gemini Pro:

The highest amount this glass measures is 600ml.



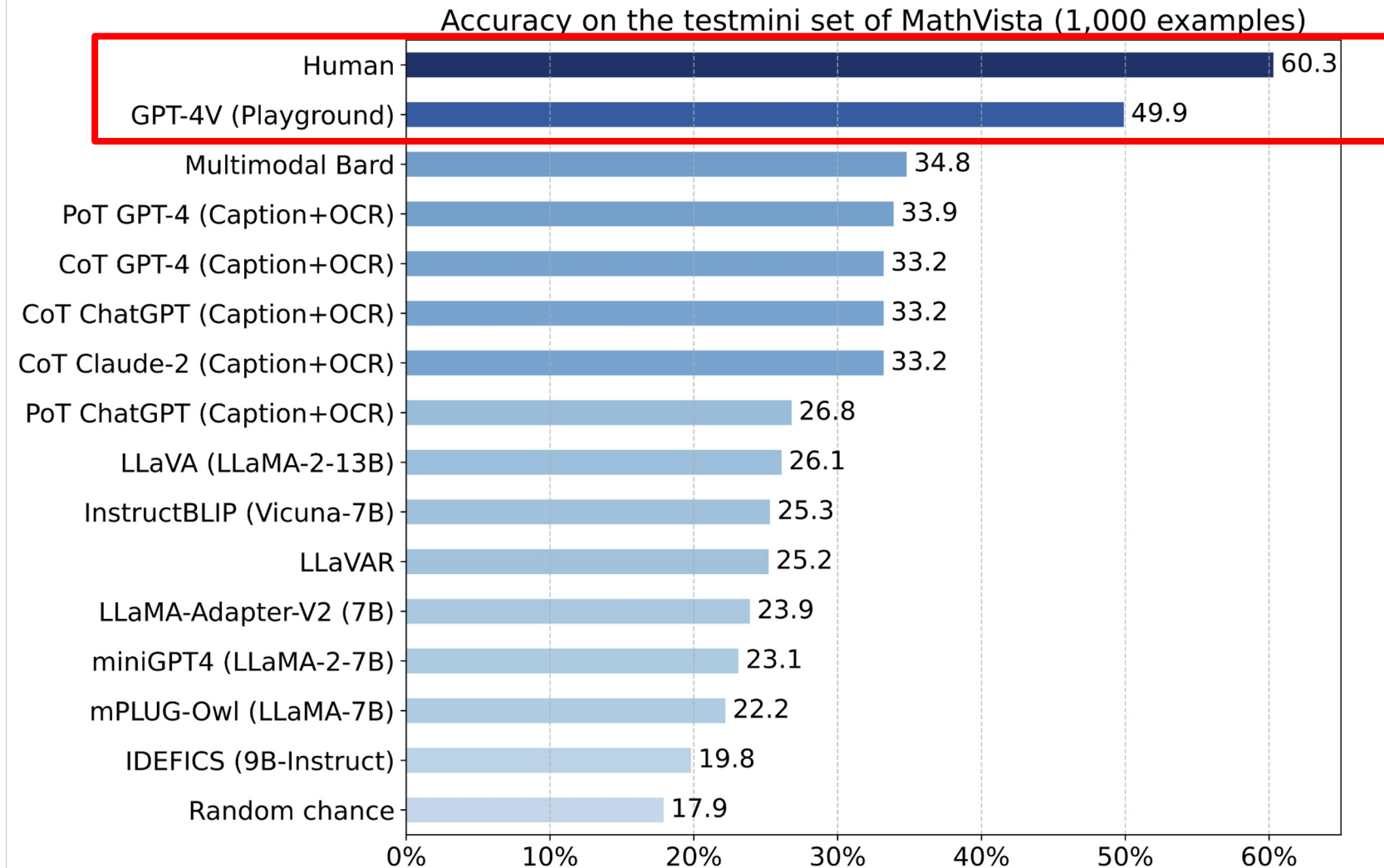
OpenAI GPT-4V

The highest amount this glass measures is 600 ml.





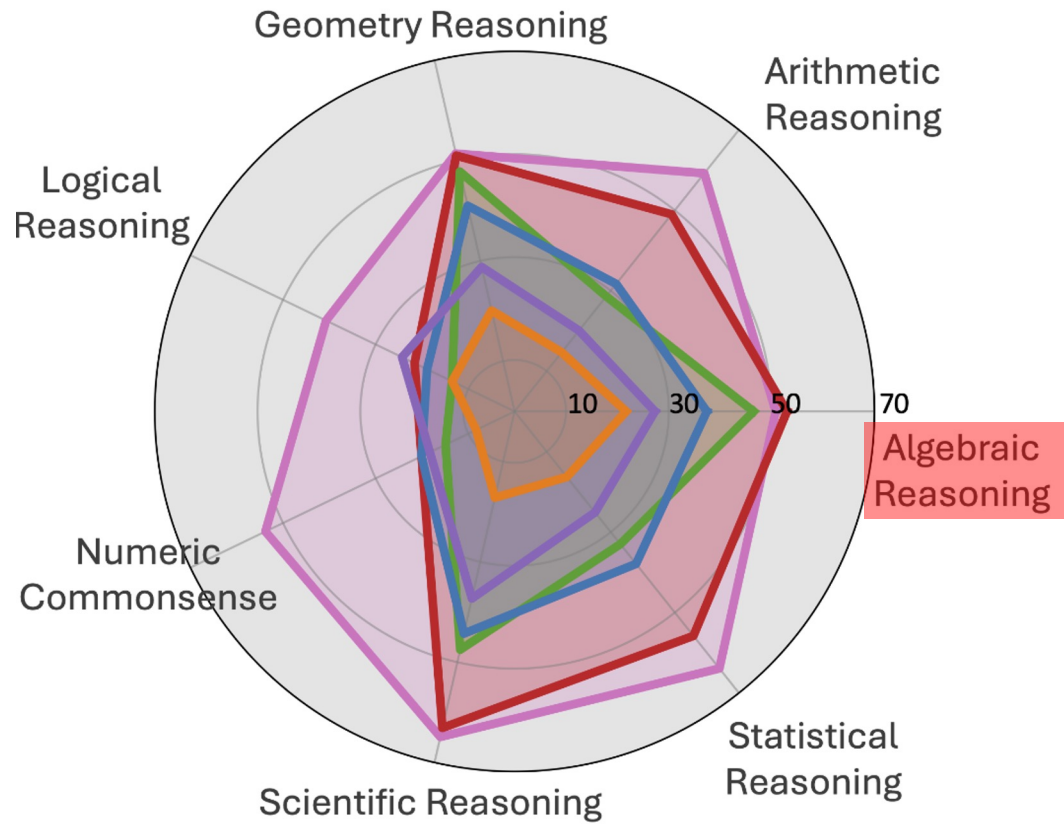
GPT-4V Falls Short of Humans by 10.4%



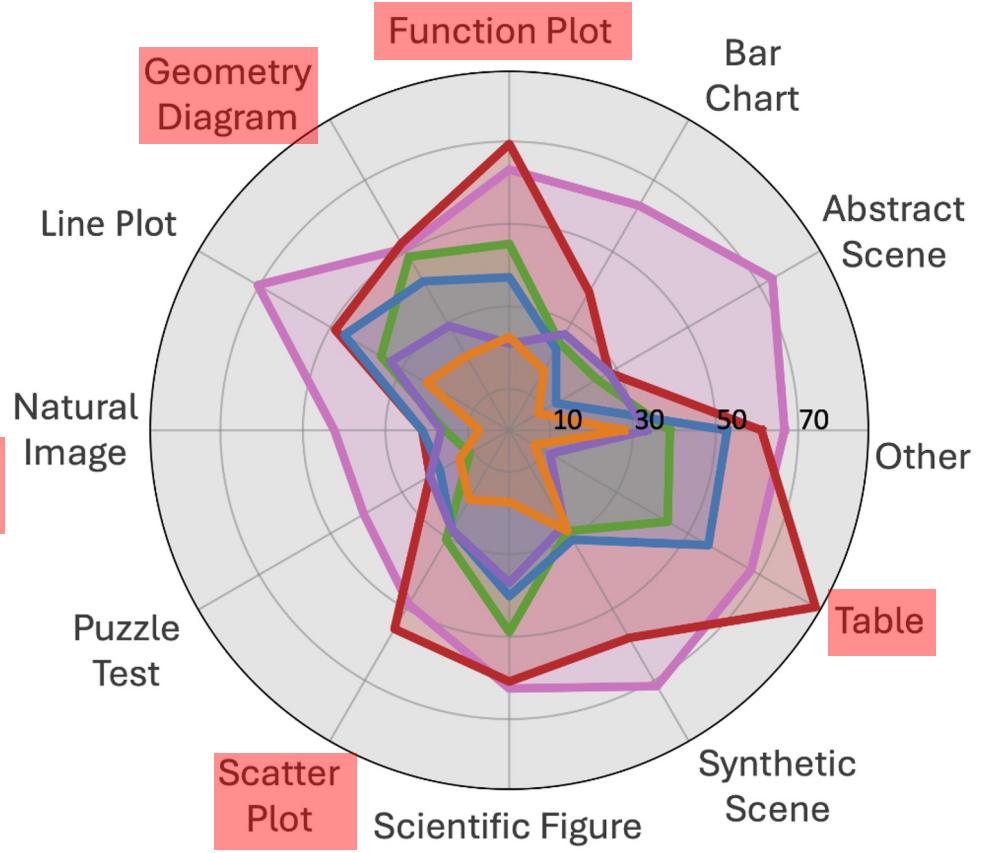


GPT-4V Outperforms Humans in Some Areas!

— Random Chance — LLaVA — PoT GPT-4 — Multimodal Bard — GPT-4V (Playground) — Human



(a) Mathematical reasoning



(b) Visual context

Tool-Augmented GPT-4 on Par with Bard

| 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)

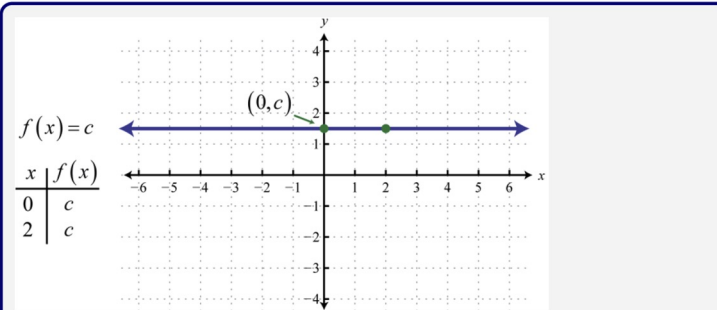
Response from GPT-4:

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

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

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.

Output: (A) polynomial

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

| Accuracy scores on the testmini s | | | | | |
|-----------------------------------|----------------------------------|---------|----------------------|------------|-------------|
| # | Model | Method | Source | Date | ALL ▾ |
| - | Human | - | Link | 2023-10-03 | 60.3 |
| 1 | GPT-4V (Playground) 🏆 | LMM 🗺️ | Link | 2023-10-15 | 49.9 |
| 2 | Multimodal Bard 🏆 | LMM 🗺️ | Link | 2023-10-03 | 34.8 |
| 3 | PoT GPT-4 (Caption+OCR) 🏆 | Tool 🗡️ | Link | 2023-10-03 | 33.9 |
| 4 | CoT GPT-4 (Caption+OCR) | Tool 🗡️ | Link | 2023-10-03 | 33.2 |
| 5 | CoT ChatGPT (Caption+OCR) | Tool 🗡️ | Link | 2023-10-03 | 33.2 |
| 6 | CoT Claude-2 (Caption+OCR) | Tool 🗡️ | Link | 2023-10-03 | 33.2 |
| 7 | PoT ChatGPT (Caption+OCR) | Tool 🗡️ | Link | 2023-10-03 | 26.8 |
| 8 | LLaVA (LLaMA-2-13B) | LMM 🗺️ | Link | 2023-10-03 | 26.1 |
| 9 | InstructBLIP (Vicuna-7B) | LMM 🗺️ | Link | 2023-10-03 | 25.3 |
| 10 | LLaVAR | LMM 🗺️ | Link | 2023-10-03 | 25.2 |
| 11 | LLaMA-Adapter-V2 (7B) | LMM 🗺️ | Link | 2023-10-03 | 23.9 |

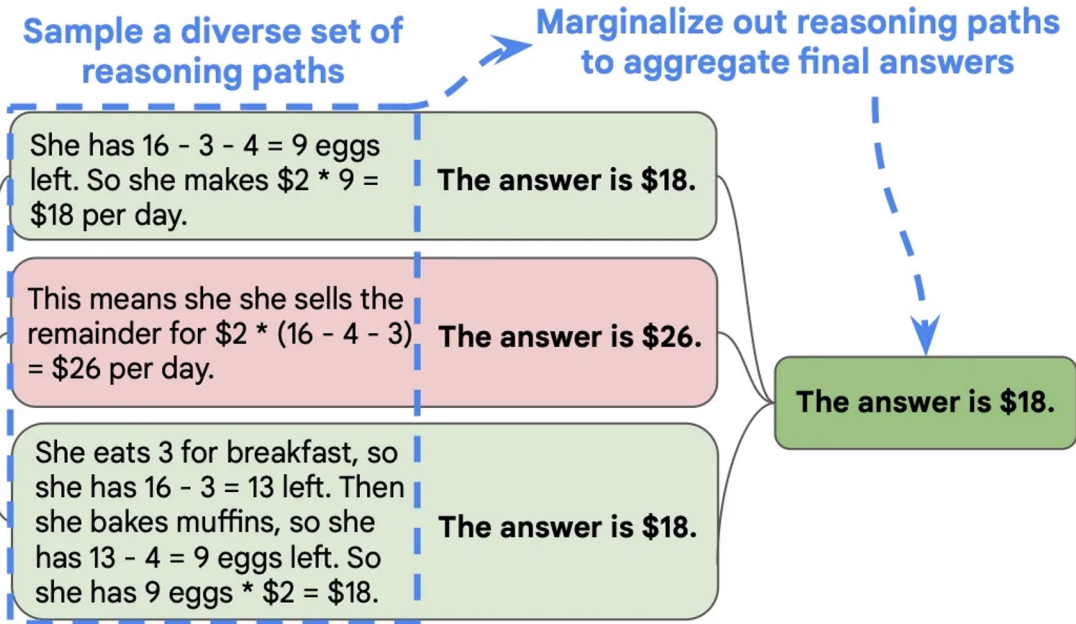


Key Finding: Self-Consistency in GPT-4V

Self-consistency

Q: *If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?*
A: There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars. The answer is 5.
...
Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for \$2 per egg. How much does she make every day?
A:

Language model





Key Finding: Self-Consistency in 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): $\text{Mean} = 512 \div 8 = 64$

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:

$$\begin{aligned}(84 + 56) &= 140 \\ (70 + 68) &= 138 \\ (60 + 47) &= 107 \\ (75 + 52) &= 127\end{aligned}$$

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

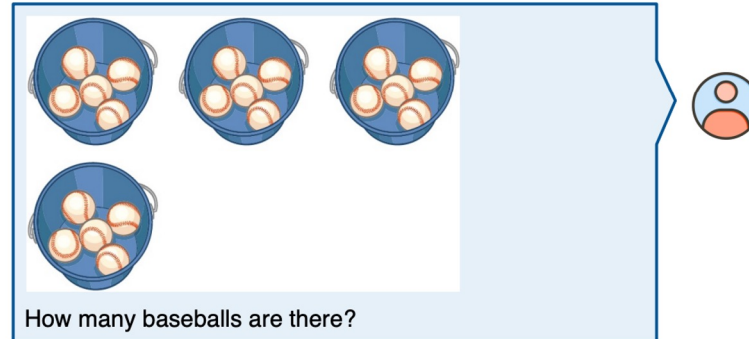
Now, I'll divide the sum by the number of individuals to get the mean: $512 \div 8 = 64$

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

Correct output: 64



GPT-4V for Multi-Turn Human-AI Interactions



Let's count the baseballs in each bucket:

1. First bucket: 5 baseballs
2. Second bucket: 6 baseballs
3. Third bucket: 6 baseballs
4. Fourth bucket: 6 baseballs

Total baseballs: $5 + 6 + 6 + 6 = 23$ baseballs.

Answer: 23.

Your solution is wrong. You counted the baseballs in each bucket incorrectly.



I apologize for the oversight. Let's try counting again:

1. First bucket: 5 baseballs
2. Second bucket: 5 baseballs
3. Third bucket: 5 baseballs
4. Fourth bucket: 5 baseballs

Total baseballs: $5 + 5 + 5 + 5 = 20$ baseballs.

Answer: 20.



GPT-4V for Multi-Turn Human-AI Interactions



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.



GPT-4V for Multi-Turn Human-AI Interactions



Sample A
Mass of each particle: 44 u
Average particle speed: 1,400 m/s

Sample B
Mass of each particle: 46 u
Average particle speed: 1,400 m/s

The diagrams below show two pure samples of gas in identical closed, rigid containers. Each colored ball represents one gas particle. Both samples have the same number of particles. Compare the average kinetic energies of the particles in each sample. Which sample has the higher temperature?
Choices: (A) sample A (B) sample B (C) neither; the samples have the same temperature



The average kinetic energy of an ideal gas particle is directly proportional to the temperature of the gas and is given by the equation:

$$KE_{avg} = \frac{3}{2} k T$$

Where:

- KE_{avg} is the average kinetic energy per molecule.
- k is the Boltzmann constant.
- T is the absolute temperature.

For gases at the same temperature and under the same conditions, the average kinetic energy is the same, irrespective of the mass of the particles.

Given that the average particle speed in both Sample A and Sample B is 1,400 m/s, and since they are in identical closed, rigid containers, it implies that both samples have the same average kinetic energy. Therefore, they also have the same temperature.

The correct answer is: (C) neither; the samples have the same temperature.



Your solution is incorrect. I am providing retrieved knowledge to help you:

The temperature of a substance depends on the average kinetic energy of the particles in the substance. The higher the average kinetic energy of the particles, the higher the temperature of the substance. **The kinetic energy of a particle is determined by its mass and speed.** For a pure substance, **the greater the mass of each particle in the substance and the higher the average speed of the particles, the higher their average kinetic energy.**



I apologize for the oversight. Let's reconsider the question.

The kinetic energy (KE) of a particle is given by the equation:

$$KE = \frac{1}{2} mv^2$$

Where:

- m is the mass of the particle.
- v is the velocity (or speed) of the particle.

Given:

- Both Sample A and Sample B have particles with the same average speed (1,400 m/s).
- The mass of each particle in Sample A is 44 u and in Sample B is 46 u.**

Given that the speeds are the same but **the mass of the particles in Sample B is greater than that of Sample A, the particles in Sample B will have a higher kinetic energy than those in Sample A.** As the temperature is directly related to the average kinetic energy of the particles, **Sample B will have a higher temperature than Sample A.**

Therefore, the correct answer is: (B) sample B.



MathVista: Among the Most Popular VQA Datasets

<https://huggingface.co/datasets/AI4Math/MathVista>

The screenshot shows the Hugging Face dataset page for AI4Math/MathVista. The page includes a search bar, a navigation menu, and a list of filters. The dataset is categorized under 'Visual Question Answering' and has 78 likes. It is available in English, Chinese, and Persian, and is monolingual. The dataset is expert-generated and found. It has 4,217 downloads last month. The page also shows the dataset viewer and a button to use the dataset in the library.

Downloads last month 4,217

The screenshot shows the Hugging Face datasets search results page. The search results are filtered by 'visual-question-answering'. The top three results are:

- MMMU/MMMU (Updated 14 days ago, 128k downloads, 106 likes)
- jmhessel/newyorker_caption_contest (Updated Dec 22, 2023, 5.22k downloads, 47 likes)
- AI4Math/MathVista** (Updated Feb 11, 4.22k downloads, 78 likes)

The AI4Math/MathVista dataset is highlighted with a red box, indicating it is one of the top 3 most downloaded VQA datasets.

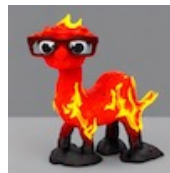


4K+ downloads per month

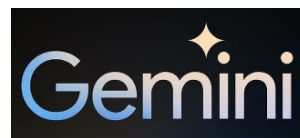


Top 3 most downloaded VQA datasets

MathVista: Advance the Latest AI Models



LLaVA 1.0, Oct 2023
LLaVA 1.6, Jan 2024



Gemini 1.0, Dec 2023
Gemini 1.5, Mar 2024



Claude 3, Mar 2024



MM1, Mar 2024



Grok-1.5V, April 2024

The Road Ahead Remains Long



MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

Brandon McKinzie[°], Zhe Gan[°], Jean-Philippe Fauconnier^{*}, Sam Dodge^{*}, Bowen Zhang^{*}, Philipp Dufter^{*}, Dhruti Shah^{*}, Xianzhi Du^{*}, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Gräsch^{*}, Alexander Toshev[†], and Yinfei Yang[†]

Apple

bmckinzie@apple.com, zhe.gan@apple.com

[°]First authors; ^{*}Core authors; [†]Senior authors



| | Claude 3 Opus | Claude 3 Sonnet | Claude 3 Haiku | GPT-4V | Gemini 1.0 Ultra |
|--|----------------------------|----------------------------|----------------------------|----------------------------|------------------|
| Math & reasoning <i>MMMU (val)</i> | 59.4% | 53.1% | 50.2% | 56.8% | 59.4% |
| Document visual Q&A <i>ANLS score, test</i> | 89.3% | 89.5% | 88.8% | 88.4% | 90.9% |
| Math <i>MathVista (testmini)</i> | 50.5% CoT | 47.9% CoT | 46.4% CoT | 49.9% | 53.0% |
| Science diagrams <i>AI2D, test</i> | 88.1% | 88.7% | 86.7% | 78.2% | 79.5% |
| Chart Q&A <i>Relaxed accuracy (test)</i> | 80.8% 0-shot CoT | 81.1% 0-shot CoT | 81.7% 0-shot CoT | 78.5% 4-shot CoT | 80.8% |



Thank You!



Evaluating Math Reasoning in Visual Contexts

Pan Lu¹, Hritik Bansal¹, Tony Xia¹, Jiacheng Liu², Chunyuan Li³, Hannaneh Hajishirzi², Hao Cheng³,
Kai-Wei Chang¹, Michel Galley³, Jianfeng Gao³

¹University of California, Los Angeles,

²University of Washington, ³Microsoft Research

ICLR 2024 Oral (85 in 7304, 1.2%)

Paper

arXiv

Code

Dataset

Visualize

Leaderboard

Twitter

<https://mathvista.github.io/>