



Stanford



UNIVERSITY OF  
NOTRE DAME

Yale

W

UNIVERSITY of  
WASHINGTON



PennState

# Benchmarking Multimodal Math Reasoning

## (Tutorial Part II)

Pan Lu

Stanford University

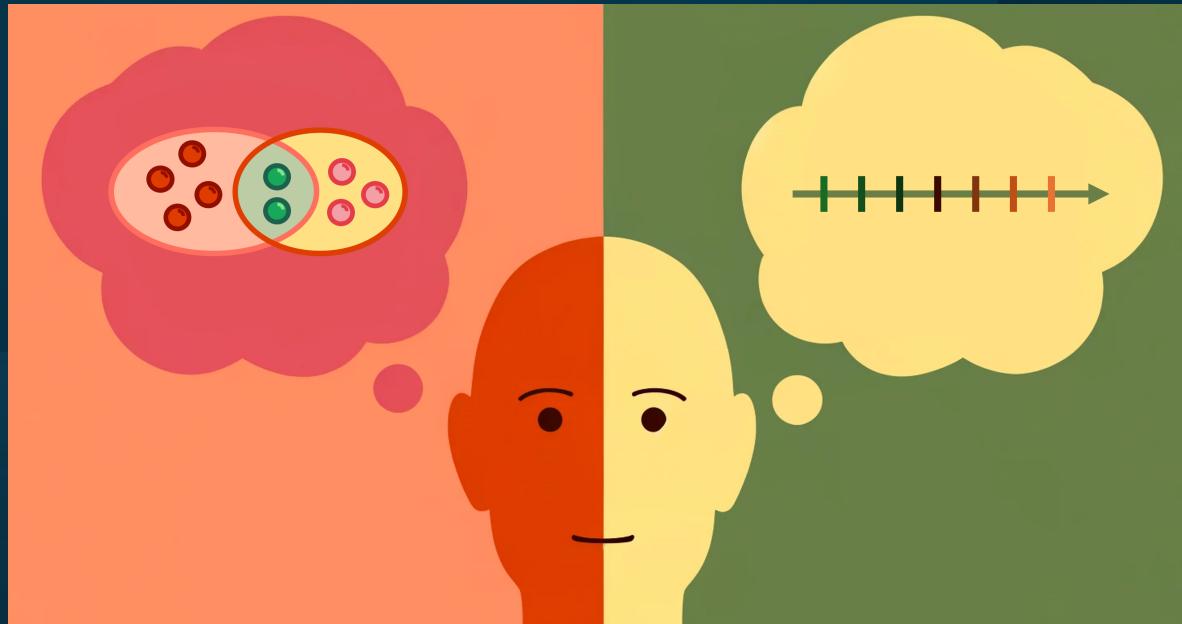
<https://lupantech.github.io/>

# Outline

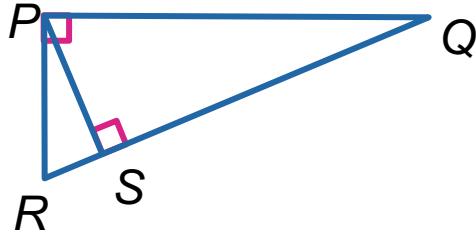
---

- What is Multimodal Math Reasoning?
- Landscape of Multimodal Math Reasoning Tasks
- Core Components of Benchmarking
- Current State-of-the-Art & Challenges
- Future Directions

# What is Multimodal Math Reasoning?



# How We Solve These Problems?



In  $\Delta PQR$ ,  $RS = 3$  and  $QS = 14$ .

Find  $PS$ .

## Understand diagram

$PS \perp RQ$ ,  
 $RP \perp PQ$ ,  
 $PS$  intersects with  $RQ$  at  $S$

## Retrieve the theorem

Pythagorean Theorem

$$PR^2 = PS^2 + RS^2$$

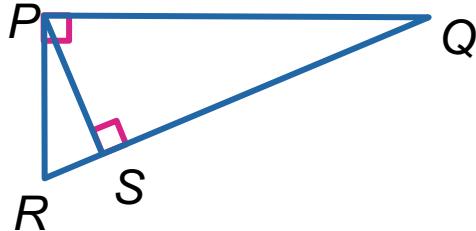
$$PQ^2 = PS^2 + QS^2$$

$$RQ^2 = RP^2 + QP^2$$



A system of quadratic equations

# How We Solve These Problems?



In  $\triangle PQR$ ,  $RS = 3$  and  $QS = 14$ .  
Find  $PS$ .

## Understand diagram

$PS \perp RQ$ ,  
 $RP \perp PQ$ ,  
 $PS$  intersects with  $RQ$  at  $S$

## Retrieve the theorem

Geometric Mean Theorem  
 $PS^2 = RS \cdot SQ$      $\frac{RS}{PS} = \frac{PS}{QS}$

## Reason (Calculate) step by step

$$\frac{RS}{PS} = \frac{PS}{QS}$$

Geometric Mean Theorem

$$\frac{3}{x} = \frac{x}{14}$$

$RS = 3$ ,  $QS = 14$ , and  $PS = x$

$$x^2 = 42$$

Cross products

$$x \approx 6.5$$

Use a calculator to take the positive square root



# How We Solve These Problems?



Miles hiked	
Day	Miles
Sunday	10
Monday	9
Tuesday	10
Wednesday	5
Thursday	9

Wanda went on a camping trip and logged the number of miles she hiked each day. What is the **median** of the numbers?

## Understand table

This table shows the number of miles Wanda hiked each day on her trip.

## Recall knowledge

To find the **median**, the data must be arranged in order from least to greatest (or greatest to least), and then the middle value(s) is/are determined.

Bus schedule		
Location	Arrive	Depart
stadium	10:20 A.M.	10:25 A.M.
park	10:35 A.M.	10:45 A.M.
hotel	11:10 A.M.	11:15 A.M.
airport	12:05 P.M.	12:10 P.M.
train station	12:25 P.M.	12:35 P.M.
bus station	1:10 P.M.	1:10 P.M.

Look at the following schedule. When does the bus **depart** from the **train station**?  
(A) 12:35 P.M. (B) 1:10 P.M.  
(C) 1:10 P.M. (D) 10:45 A.M.

## Understand table

This table shows the schedule of the bus. There are multiple rows and columns

## Reason (Program) step by step

Call external tool (Python Interpreter)

```
miles = [10, 9, 10, 5, 9]
miles= sorted(miles)
mid1 = (len(miles) - 1) // 2
mid2 = len(miles) // 2
ans = (miles[mid1] + miles[mid2]) / 2
# executed output: ans = 9
```

## Reason (Think) step by step

Call external tools (row lookup)

Location	Arrive	Depart
train station	12:25 P.M.	12:35 P.M.

- Find the train station on the schedule. Find the departure time for the train station.
- Train station: 12:35 P.M. The bus departs from the train station at 12:35 P.M.
- The answer is **12:35 P.M.**

# How We Solve These Problems?



**Context:** A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.



**Question:** Which type of force from the baby's hand opens the cabinet door?

**Options:** (A) pull (B) push

## Understand multimodal context

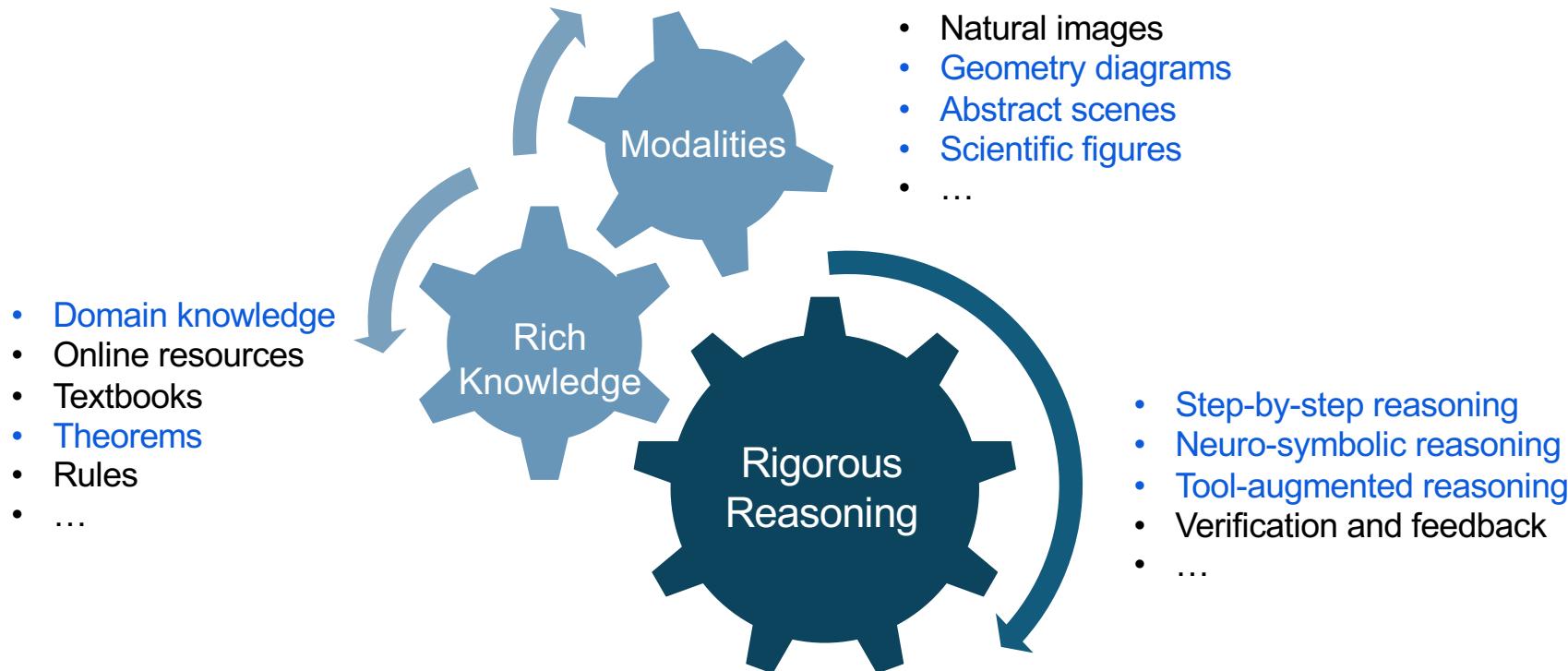
A baby opens the cabinet door.

## Recall background knowledge

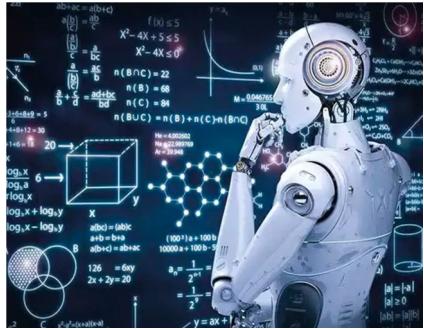
The direction of a push is away from the object that is pushing. The direction of a pull is toward the object that is pulling.

## Reason (Think) step by step

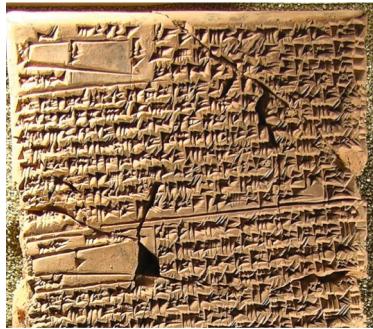
1. The baby's hand applies a force to the cabinet door.
2. This force causes the door to open.
3. The direction of this force is toward the baby's hand.
4. This force is a pull.



# Why Multimodal Reasoning in Math Matters?



Fundamental AI Problem



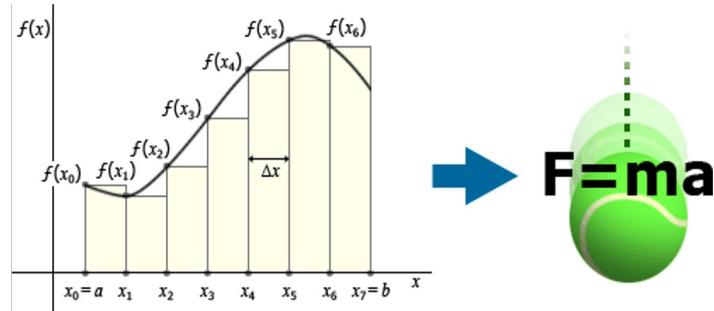
2000BC, Mesopotamia



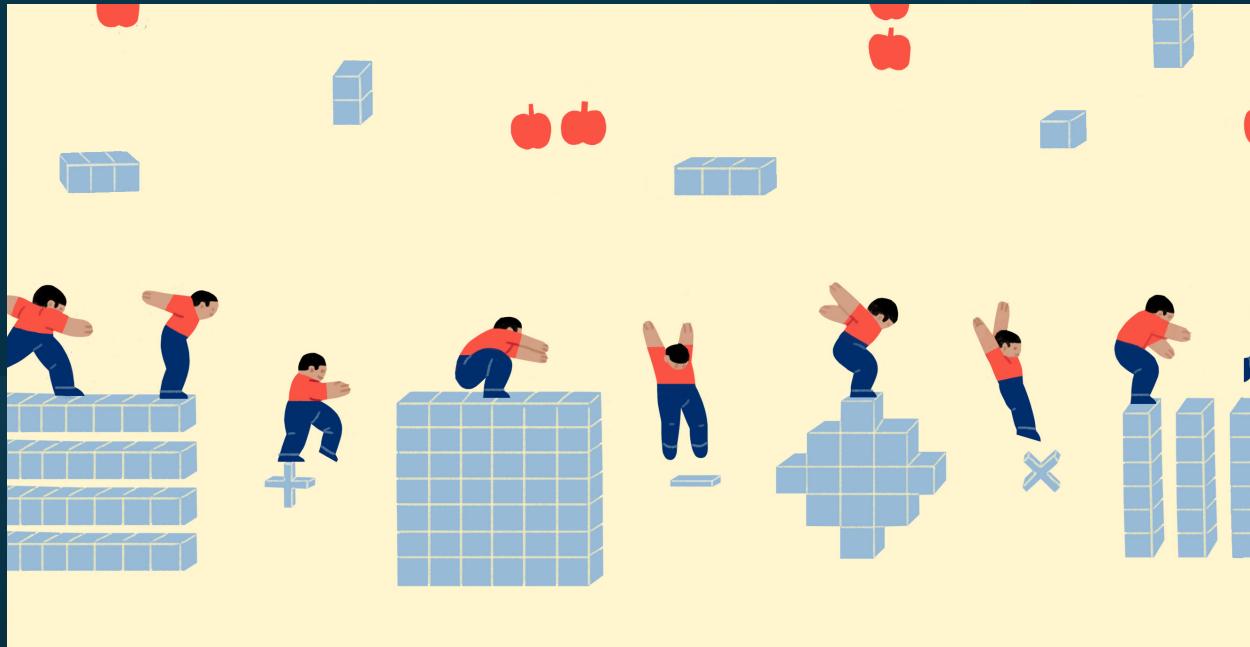
K-12 Education



Wide Applications



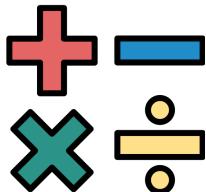
# Landscape of Multimodal Math Reasoning Tasks



# Scope: What are we trying to solve?



## 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:  $\$9.08 \times 4 = \$36.32$

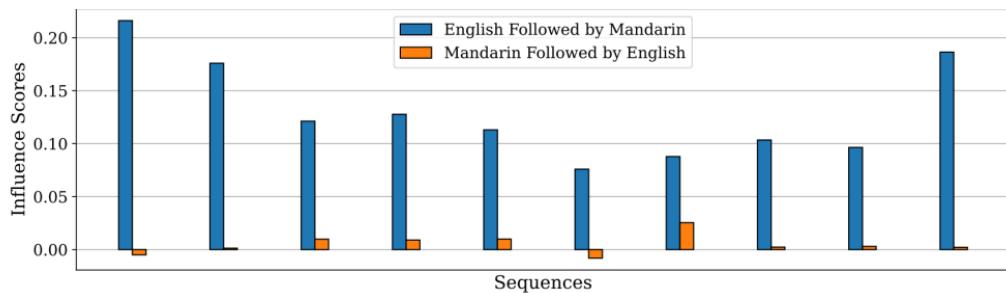
Find the cost of the canvas scraps. Multiply:  $\$8.17 \times 4 = \$32.68$

Now find the total cost by adding:  $\$36.32 + \$32.68 = \$69$

She spent \$69.

**Answer:** 69

# Scope: What are we trying to solve?



Statistical

**Question:** How many sequences have negative Influence Scores?

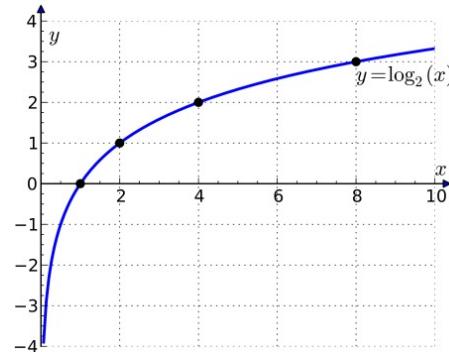
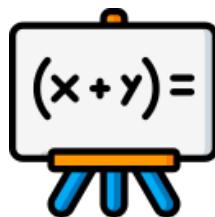
**Answer:** 2



# Scope: What are we trying to solve?



Algebraic



**Question:** The derivative of  $y$  at  $x = 6$  is \_\_\_\_ than at  $x = 8$ .

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

**Answer:** (A) larger than

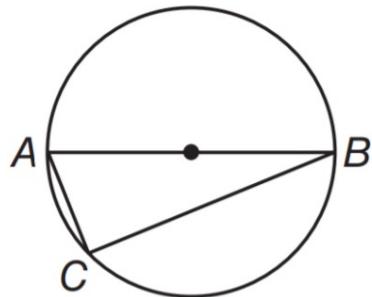
**Question:** How many zeros does this function have?

**Answer:** 1

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

**Answer:** 0

# Scope: What are we trying to solve?



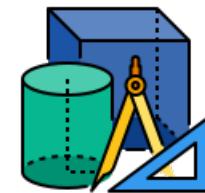
**Question:**  $\overline{AB}$  is a diameter,  $AC = 8$  inches, and  $BC = 15$  inches. Find the radius of the circle.

**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

Geometry



# Scope: What are we trying to solve?

Numeric



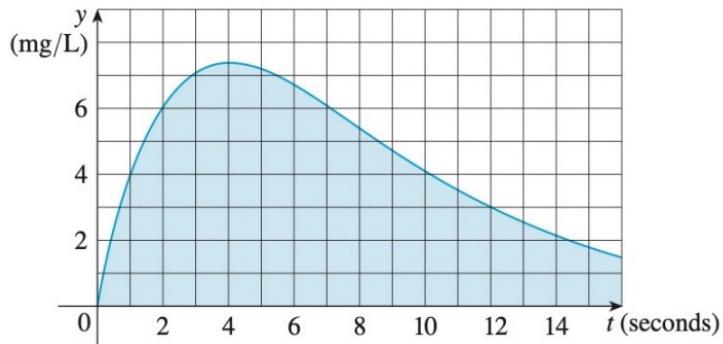
**Question:** What is the age gap between these two people in image? (unit: years)

**Named entities:** Winston Churchill, Charles de Gaulle

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

**Answer:** 16

# Scope: What are we trying to solve?



**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

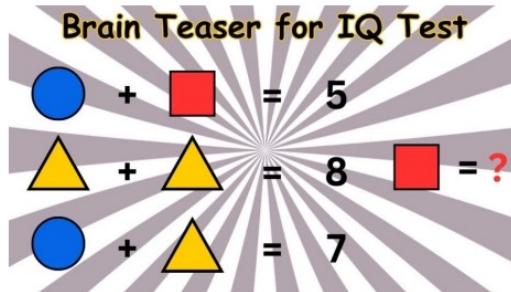
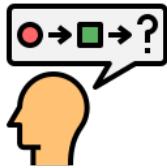
Scientific



# Scope: What are we trying to solve?



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.

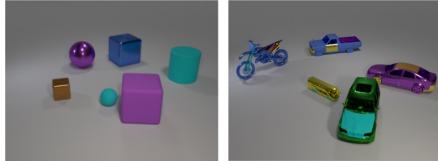
Therefore Square = 2

**Answer:** 2

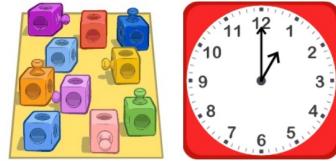
# Modalities: What are we trying to solve?



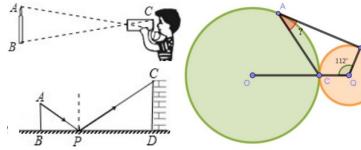
Natural Images



Synthetic Scene



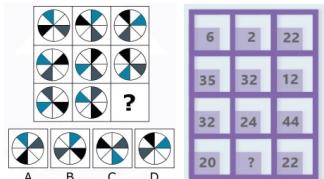
Abstract Scene



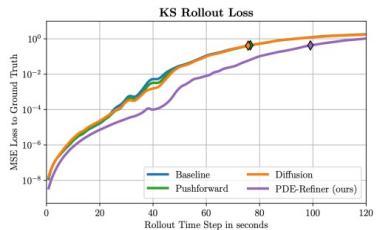
Geometry Diagram

Cans of food collected	
Name	Number of cans of food
Emmett	8
Luther	7
Bruce	10
Scott	9
Mabel	9
Roxanne	5
Kevin	8
Pluto	900
	248
	2.99

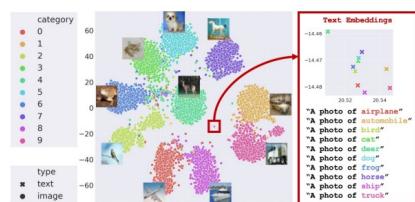
Table



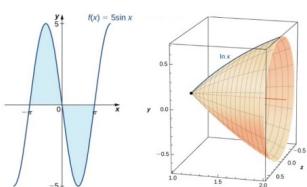
Puzzle Test



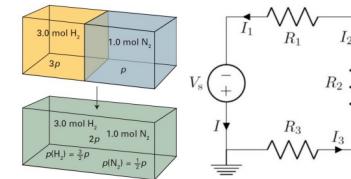
Line Plot



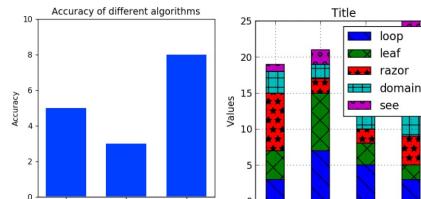
Bar Chart



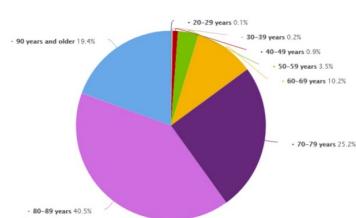
Function Plot



Scientific Figure



Scatter Plot

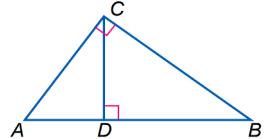


Pie Chart

# Core Components of Benchmarking



# Geometry3K: Geometric Reasoning



**Question:**

In triangle ABC, AD = 3 and BD = 14. Find CD.

**Choices:**

- (A) 6.0 (B) 6.5 (C) 7.0 (D) 8.5  
Answer: (B) 6.5

**Diagram Formal Language**

```
Triangle(A,B,C)
Triangle(A,C,D)
Triangle(B,C,D)
PointLiesOnLine(D,Line(A,B))
Perpendicular(Line(A,C),Line(B,C))
Perpendicular(Line(C,D),Line(A,B))
```

**Text Formal Language**

```
Triangle(A,B,C)
Equals(LengthOf(Line(A,D)),3)
Equals(LengthOf(Line(B,D)),14)
Find(LengthOf(Line(C,D)))
```

## Formal Annotations

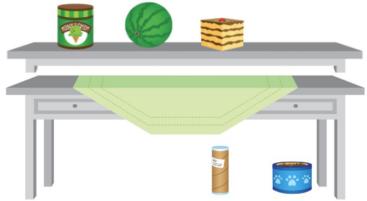
Problem Text	Diagram	Choices	Text Literals (Logic Forms)	Diagram Literals (Logic Forms)
In triangle ABC, AD = 3 and BD = 14. Find CD.		A. 6.0 B. 6.5 C. 7.0 D. 8.5 Answer: B	Triangle(A,B,C) Equals(LengthOf(Line(A,D)),3) Equals(LengthOf(Line(B,D)),14) Find(LengthOf(Line(C,D)))	Triangle(A,B,C) Triangle(A,C,D) Triangle(B,C,D) PointLiesOnLine(D,Line(A,B)) Perpendicular(Line(A,C),Line(B,C)) Perpendicular(Line(C,D),Line(A,B))
Find y. Round to the nearest tenth.		A. 18.8 B. 23.2 C. 25.9 D. 44.0 Answer: C	Find(y)	Equals(LengthOf(Line(A,B)),32) Equals(LengthOf(Line(B,D)),y) Equals(MeasureOf(Angle(A,C,B)),54) Equals(LengthOf(Line(A,D)),x) PointLiesOnLine(D,Line(A,C)) Perpendicular(Line(B,D),Line(C,D)) Equals(LengthOf(Line(A,B)),LengthOf(Line(B,C)))
Find the perimeter of \$parallelogram\$ JKLM.		A. 11.2 B. 22.4 C. 24 D. 44.8 Answer: B	Find(PerimeterOf(Parallelogram(J,K,L,M)))	Equals(LengthOf(Line(L,K)),7.2) Equals(LengthOf(Line(M,L)),4) Equals(LengthOf(Line(E,J)),6) PointLiesOnLine(E,Line(M,L)) Perpendicular(Line(J,E),Line(E,L))
In \$\odot K\$, MN = 16 and m \$\widehat{MN} = 98\$. Find the measure of LN. Round to the nearest hundredth.		A. 6.93 B. 7.50 C. 8.94 D. 10.00 Answer: C	Circle(K) Equals(LengthOf(Line(M,N)),16) Equals(MeasureOf(Arc(M,N)),98) Find(LengthOf(Line(L,N)))	Equals(LengthOf(Line(J,K)),10) Perpendicular(Line(P,K),Line(M,P)) PointLiesOnLine(P,Line(M,N)) PointLiesOnLine(P,Line(L,J)) PointLiesOnLine(P,Line(L,K)) PointLiesOnLine(K,Line(P,J)) PointLiesOnLine(K,Line(L,J)) PointLiesOnCircle(M,Circle(K)) PointLiesOnCircle(J,Circle(K)) PointLiesOnCircle(N,Circle(K)) PointLiesOnCircle(L,Circle(K))

# IconQA: Abstract Diagram Reasoning



Q: What is the man doing?  
A: **riding a motorcycle**

VQA



Q: Which object is next to the one shaped like a cube?

- C:
- - 
  - 
  - 
  -



Q: How many tomatoes are there?  
A: **5**

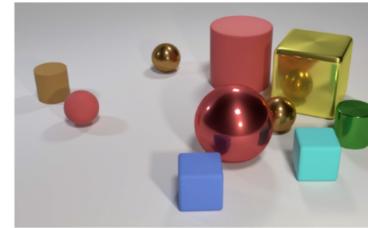
VQA 2.0



Q: Which picture shows the pizza inside the oven?

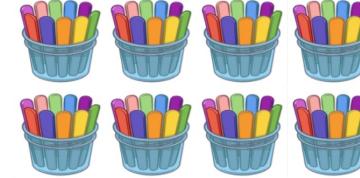
- C: **(A) left one** (B) right one

IconQA



Q: How many objects are metal things?  
A: **4**

CLEVR



Q: How many sticks are there?  
A: **80**

100K+ Diagram QAs

# TabMWP: Semi-structured Reasoning



square beads	\$2.97 per kilogram
oval beads	\$3.41 per kilogram
flower-shaped beads	\$2.18 per kilogram
star-shaped beads	\$1.95 per kilogram
heart-shaped beads	\$1.52 per kilogram
spherical beads	\$3.42 per kilogram
rectangular beads	\$1.97 per kilogram

**Question:** If Tracy buys 5 kilograms of spherical beads, 4 kilograms of star-shaped beads, and 3 kilograms of flower-shaped beads, how much will she spend? (unit: \$)

**Answer:** **31.44**

**Solution:**

Find the cost of the spherical beads. Multiply:  $\$3.42 \times 5 = \$17.10$ .

Find the cost of the star-shaped beads. Multiply:  $\$1.95 \times 4 = \$7.80$ .

Find the cost of the flower-shaped beads. Multiply:  $\$2.18 \times 3 = \$6.54$ .

Now find the total cost by adding:  $\$17.10 + \$7.80 + \$6.54 = \$31.44$ .

She will spend **31.44**.

Sandwich sales		
Shop	Tuna	Egg salad
City Cafe	6	5
Sandwich City	3	12
Express Sandwiches	7	17
Sam's Sandwich Shop	1	6
Kelly's Subs	3	4

**Question:** As part of a project for health class, Cara surveyed local delis about the kinds of sandwiches sold. Which shop sold fewer sandwiches, Sandwich City or Express Sandwiches?

**Options:** (A) Sandwich City (B) Express Sandwiches

**Answer:** **(A) Sandwich City**

**Solution:**

Add the numbers in the Sandwich City row. Then, add the numbers in the Express Sandwiches row.

Sandwich City:  $3 + 12 = 15$ . Express Sandwiches:  $7 + 17 = 24$ .

15 is less than 24. **Sandwich City** sold fewer sandwiches.

Step-wise solutions

# The MathVista Benchmark



Total Downloads ◇

150k

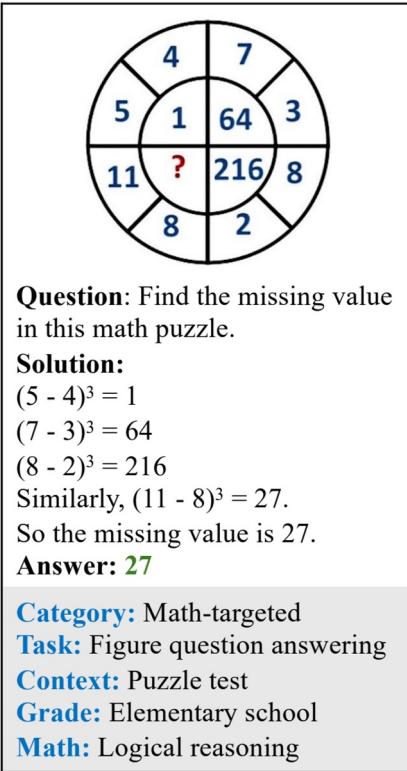
Time evolution

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

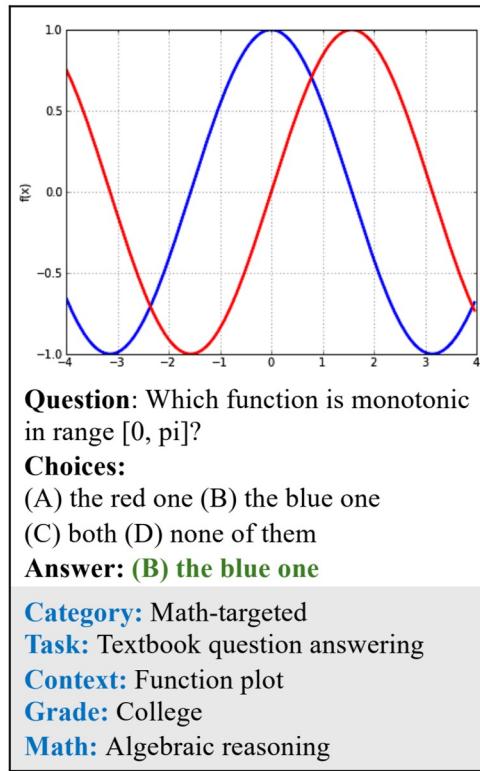
Demo: <https://mathvista.github.io/#visualization>

The screenshot shows the MathVista Visualizer interface. On the left, a modal window titled "MathVista Visualizer" contains "Sample Filters" with dropdown menus for various parameters like "How many samples?", "Choose a split:", "Choose a question type:", etc. At the bottom of this window is a blue "Refresh data!" button. To the right, there are three main sections: 1) A "Question" section with an image of a solar eclipse and choices "earth", "moon", "sun", and "solar eclipse". 2) Another "Question" section with a diagram of a rectangle LMNP divided into four triangles (LNM, MNP, LMP, NMP) and choices 2, 4, 5, 10. 3) A third "Question" section with a table of cruise ticket prices and a choice "Most preferred objects of different categories".

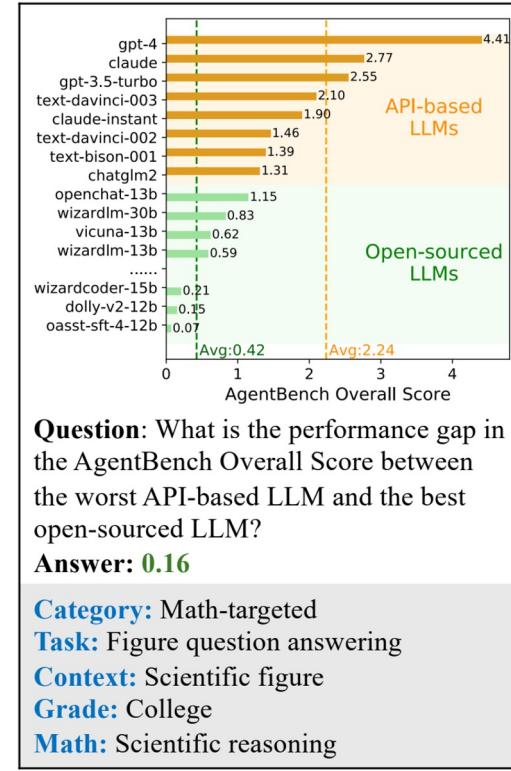
# Three New Curated Datasets



(a) IQTest



(b) FunctionQA



(c) PaperQA

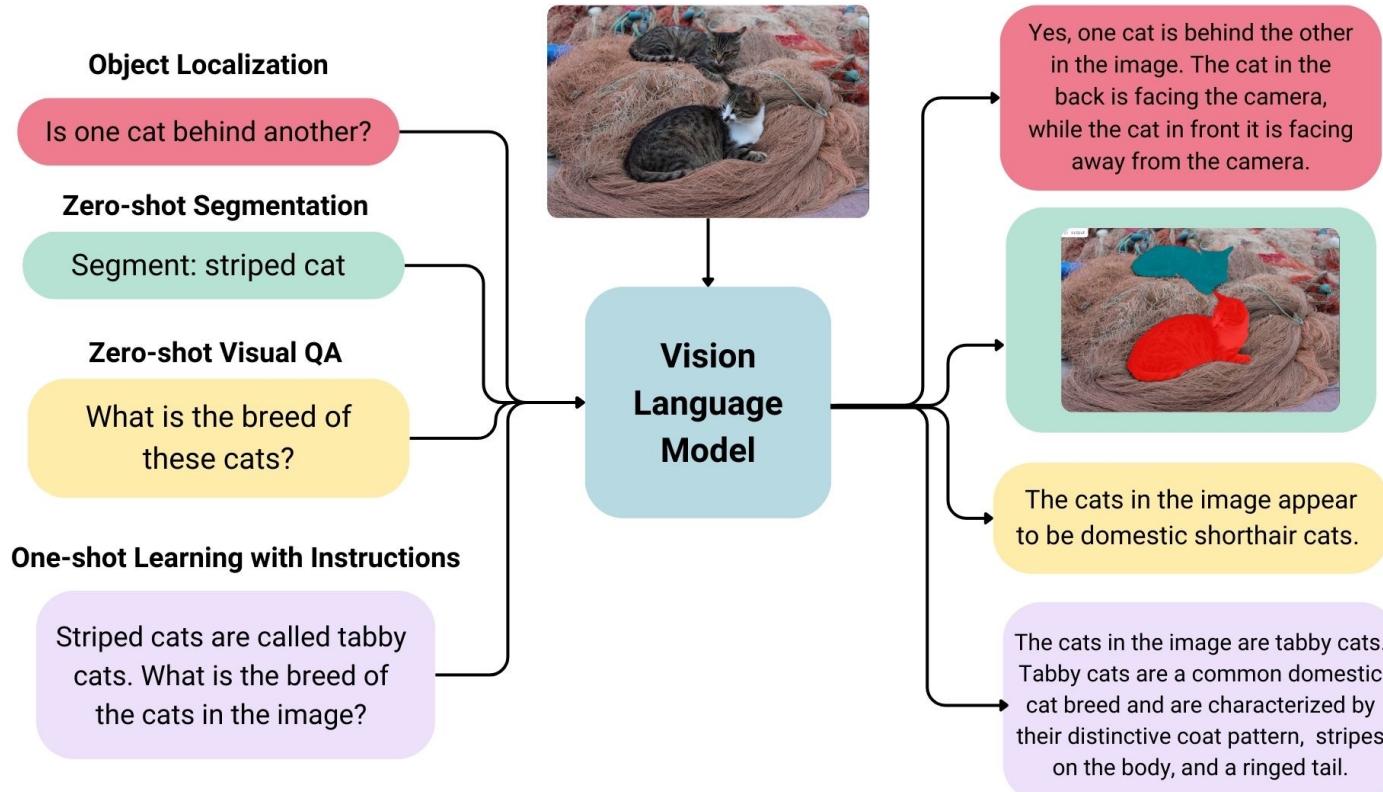
- **Response Generation**

Question type	Answer type	Task instruction
multiple-choice	Text	Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
Free-form	Integer	Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end.
Free-form	Float (1)	Please answer the question requiring a floating-point number with one decimal place and provide the final value, e.g., 1.2, 1.3, 1.4, at the end.
Free-form	Float (2)	Please answer the question requiring a floating-point number with two decimal places and provide the final value, e.g., 1.23, 1.34, 1.45, at the end.
Free-form	List	Please answer the question requiring a Python list as an answer and provide the final list, e.g., [1, 2, 3], [1.2, 1.3, 1.4], at the end.

- Answer Extraction & Score Calculation

Element	Prompt
Task description	<p>Please read the following example. Then extract the answer from the model response and type it at the end of the prompt.</p> <p><b>Hint:</b> Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end.</p> <p><b>Question:</b> Which number is missing?</p>
Example 1	<p><b>Model response:</b> The number missing in the sequence is 14.</p> <p><b>Extracted answer: 14</b></p> <p><b>Hint:</b> Please answer the question requiring a floating-point number with one decimal place and provide the final value, e.g., 1.2, 1.3, 1.4, at the end.</p> <p><b>Question:</b> What is the fraction of females facing the camera?</p>
Example 2	<p><b>Model response:</b> The fraction of females facing the camera is 0.6, which means that six out of ten females in the group are facing the camera.</p> <p><b>Extracted answer: 0.6</b></p>

# Models & Approaches Being Benchmarked



# A Geometry Problem

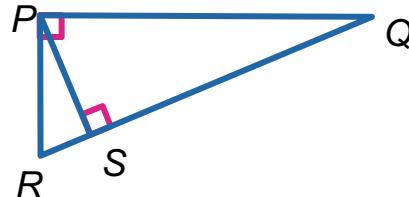
In  $\triangle PQR$ ,  $RS = 3$  and  $QS = 14$ . Find  $PS$ .

$$\frac{RS}{PS} = \frac{PS}{QS} \quad \text{Theorem 8.2}$$

$$\frac{3}{x} = \frac{x}{14} \quad RS = 3, QS = 14, \text{ and } PS = x$$

$x^2 = 42$       Cross products

$x \approx 6.5$       Use a calculator to take the positive square root of each side.



How can we represent the geometry problem in a way that is more **accessible to computers**?

# A Formal Language in Geometry



Terms	Examples
<i>predicate</i>	Line, IntersectAt, IsMedianOf
<i>literal</i>	Find(AreaOf(Triangle(A,B,C)))

Order-1 Literals

```
Triangle(A,B,C)
Triangle(C,B,D)
Line(A,D)
Line(C,D)
Line(B,C)
```

Order-2 Literals

```
PerimeterOf(Angle(A,B,C))
Similar(Triangle(A,B,C),Triangle(C,B,D))
LengthOf(Line(A,D))
LengthOf(Line(C,D))
LengthOf(Line(B,C))
```

Order-3 Literals

```
Find(PerimeterOf(Angle(A,B,C)))
Equals(LengthOf((Line(A,D)),5)
Equals(LengthOf(Line(C,D)),12)
Equals(LengthOf(Line(B,C)),31.2)
```

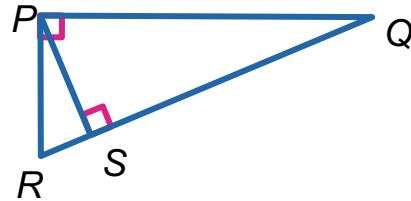
# Formal Language in Geometry



In  $\triangle PQR$ ,  $RS = 3$  and  $QS = 14$ . Find  $PS$ .

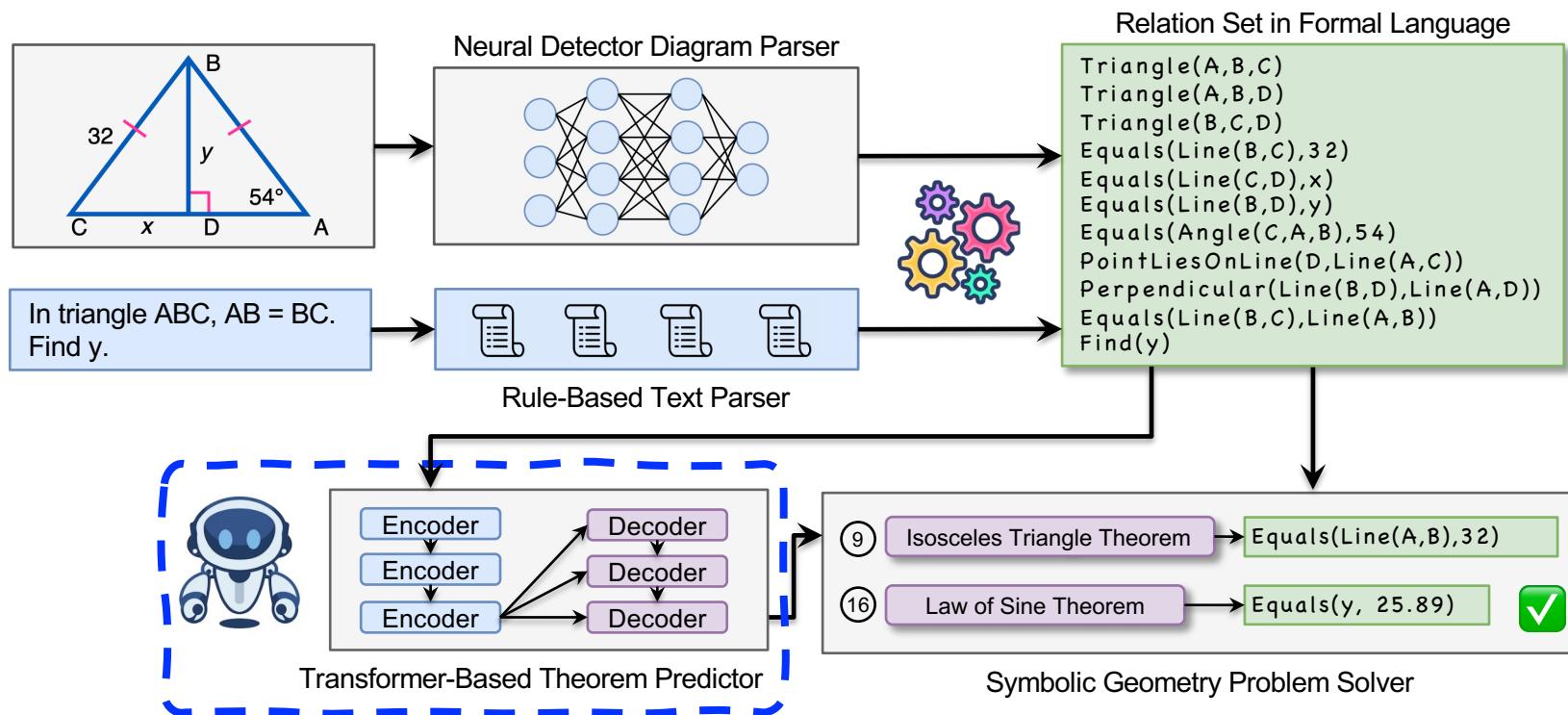


```
Triangle(P,Q,R)
Equals(LengthOf(Line(R,S)),3)
Equals(LengthOf(Line(Q,S)),14)
Find(LengthOf(Line(P,S)))
```

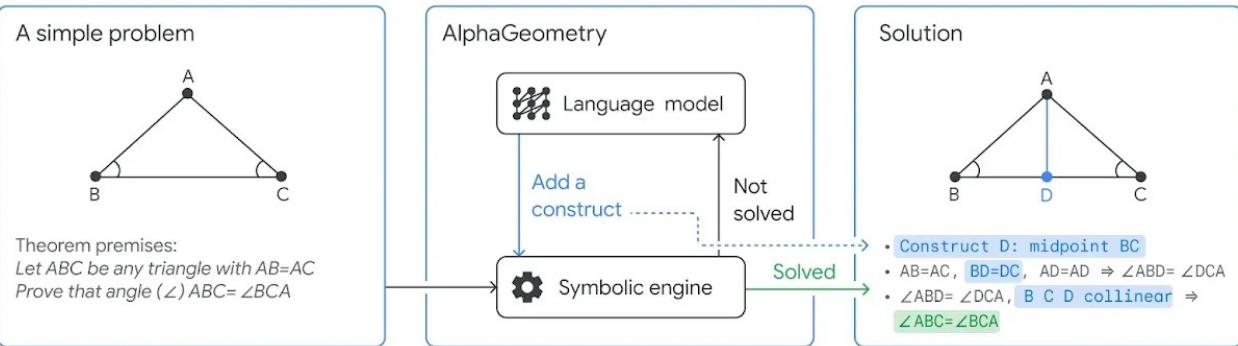


```
Triangle(P,Q,R)
Triangle(S,R,P)
Triangle(S,P,Q)
Perpendicular(Line(R,S),Line(Q,S))
Perpendicular(Line(R,P),Line(Q,P))
PointLiesOnLine(S,Line(R,Q))
```

# Inter-GPS: Neuro-Symbolic Solver



# AI System for Geometry



Google's AlphaGeometry (2024)  
**Olympiad-level** performance for the first time

<https://deepmind.google/discover/blog/alphageometry-an-olympiad-level-ai-system-for-geometry/>

nature

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 17 January 2024

**Solving olympiad geometry without human demonstrations**

Trieu H. Trinh ✉, Yuhuai Wu, Quoc V. Le, He He & Thang Luong ✉

# Can One Model Solve Diverse Problems?



**Context:** A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.



**Question:** Which type of force from the baby's hand opens the cabinet door?

**Options:** (A) pull (B) push



Miles hiked	
Day	Miles
Sunday	10
Monday	9
Tuesday	10
Wednesday	5
Thursday	9

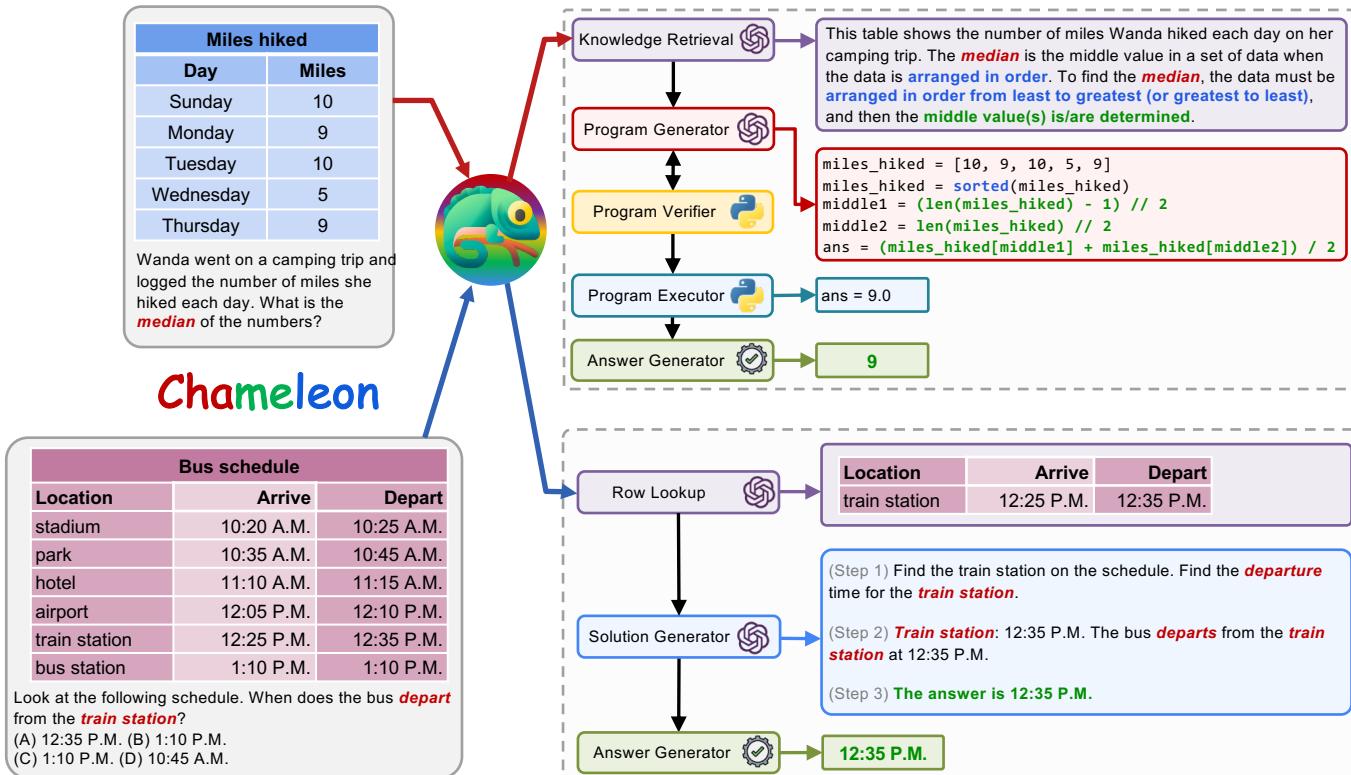
Wanda went on a camping trip and logged the number of miles she hiked each day. What is the **median** of the numbers?

Bus schedule		
Location	Arrive	Depart
stadium	10:20 A.M.	10:25 A.M.
park	10:35 A.M.	10:45 A.M.
hotel	11:10 A.M.	11:15 A.M.
airport	12:05 P.M.	12:10 P.M.
train station	12:25 P.M.	12:35 P.M.
bus station	1:10 P.M.	1:10 P.M.

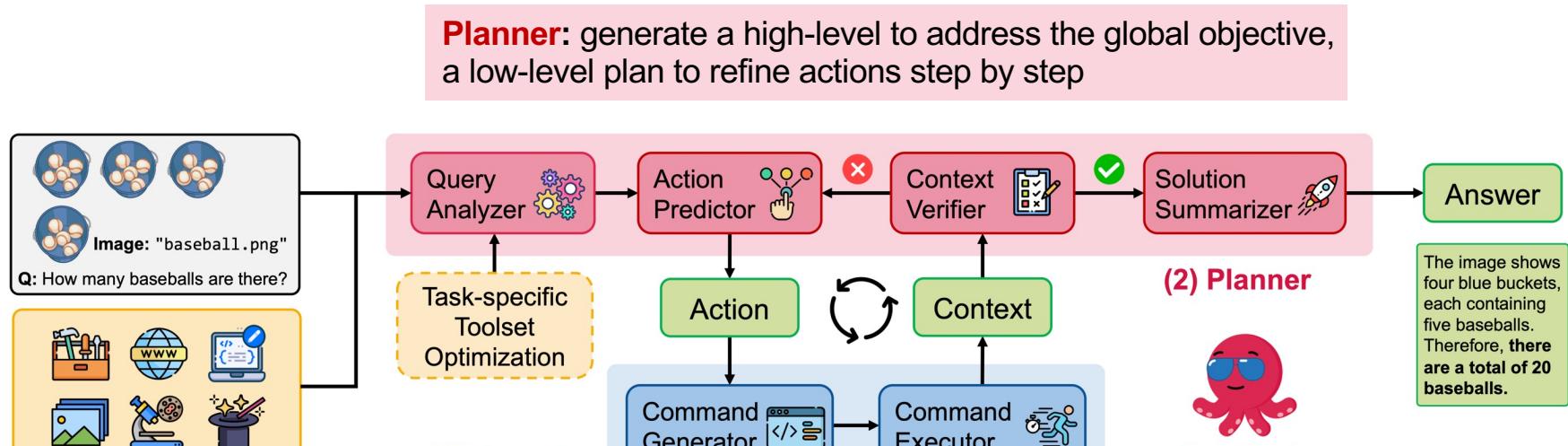
Look at the following schedule. When does the bus **depart** from the **train station**?  
 (A) 12:35 P.M. (B) 1:10 P.M.  
 (C) 1:10 P.M. (D) 10:45 A.M.

Rock collection		Watermelons harvested		Birthday party		Guests per hour	
Day	Number of rocks	Day	Number of watermelons	Activity	Parents	Children	Guests per hour
the school	9:00 A.M. 10:15 A.M. 10:30 A.M. 10:45 A.M. 11:00 A.M.	topaz	\$18.55 per lb	Singing	14	20	1, 1.5, 6
the zoo	9:00 A.M. 10:15 A.M. 10:30 A.M. 10:45 A.M. 11:00 A.M.	amethyst	\$19.88 per lb	Eating cake	8	10	2, 2.5, 8
the mall	9:00 A.M. 10:15 A.M. 10:30 A.M. 10:45 A.M. 11:00 A.M.	tiger's eye	\$19.29 per lb	Jumping rope	16	20	4, 5
the grocery store	9:00 A.M. 10:15 A.M. 10:30 A.M. 10:45 A.M. 11:00 A.M.	topaz	\$19.00 per lb	Swimming	16	19	6, 6.5, 2
the science museum	9:00 A.M. 10:15 A.M. 10:30 A.M. 10:45 A.M. 11:00 A.M.	topaz	\$18.63 per lb	Playing tag	4	9	8, 8.5, 6, 7
the library	9:00 A.M. 10:15 A.M. 10:30 A.M. 10:45 A.M. 11:00 A.M.	topaz	\$19.23 per lb				
the football field	11:00 A.M. 1:00 P.M. 1:15 P.M. 1:30 P.M. 2:45 P.M.	rectangular beads	\$2/kilogram				
the playground	12:00 P.M. 2:00 P.M. 2:15 P.M. 2:45 P.M. 3:45 P.M.	spherical beads	\$2/kilogram				
the doctor's office	1:00 P.M. 2:00 P.M. 2:45 P.M. 3:15 P.M. 4:15 P.M.	oval beads	\$2/kilogram				
		heart-shaped beads	\$3/kilogram				
		rectangular beads	\$2/kilogram				
		spherical beads	\$2/kilogram				
		oval beads	\$2/kilogram				
Tour boat schedule		Employee		Per period		Guests per hour	
Ocean City		Diana Khalid		December 6-13		\$620.00	
Whale Watch Harbor		Total earnings		\$71.14		1	
Oyster Lighthouse		Federal income tax		\$7.14		2	
Fisherman's Cove		State income tax		\$48.90		3	
Surfing Beach		Other taxes		\$47.00		4	
		Fees		7		5	
		Pay after taxes		7		6	
Option		Change in phone price		Guests per hour			
Add an upgrade		\$60					
Buy a used phone		-\$75					
Resort guest count							
Park		Private Park		2			
Kelly's Park		Public Park		4			
Premiere Park		Grand Park		4			
Adventure Park		Mystic Park		3			
Crusade Park		Crown Park		2			
Mountain Park		Mystique Park		4			

# Chameleon for Mathematical Reasoning



# The OctoTools Framework



**Tool cards:** define tool-usage metadata, encapsulate tools

**Executor:** instantiate tool calls by generating executable commands, save structured results in the context

# Current State-of-the-Art & Challenges



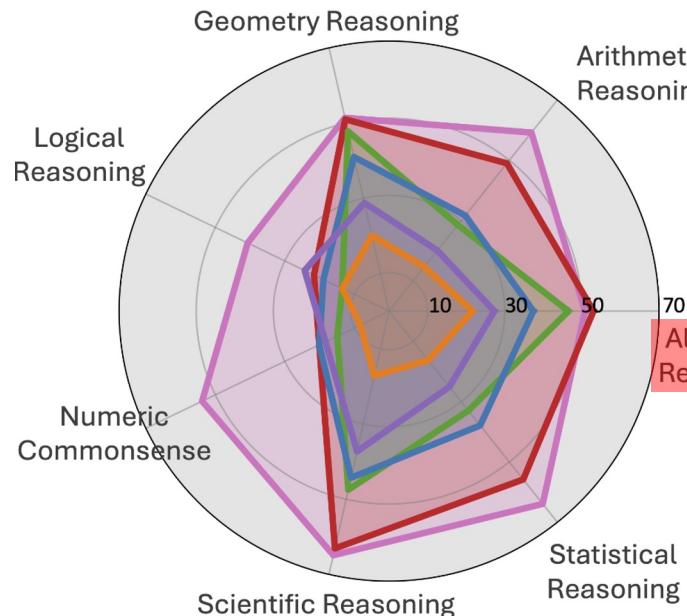
# MathVista Leaderboard



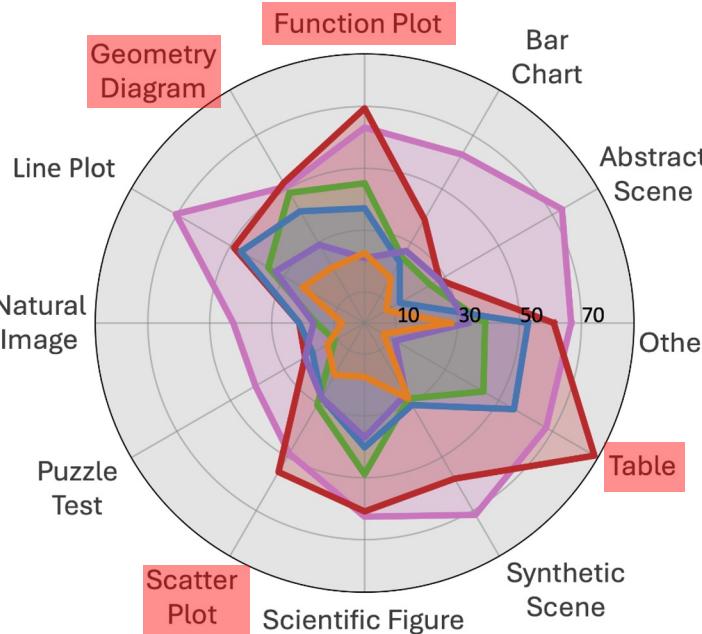
<https://mathvista.github.io/#leaderboard>

#	Model	Method	Source	Date	ALL	FQA	GPS	MWP	TQA	VQA	ALG	ARI	GEO	LOG	NUM	SCI	STA
-	Human Performance*	-	<a href="#">Link</a>	2023-10-03	60.3	59.7	48.4	73.0	63.2	55.9	50.9	59.2	51.4	40.7	53.8	64.9	63.9
1	Step R1-V-Mini 🏆	Reason 🧠	<a href="#">Link</a>	2025-04-07	80.1	80.3	89.9	87.6	77.2	63.1	86.1	75.6	87.4	54.1	56.3	77.9	87.7
2	Kimi-k1.6-preview-20250308 🌟	Reason 🧠	<a href="#">Link</a>	2025-03-10	80.0	82.9	91.8	82.3	76.0	63.1	87.2	73.1	89.1	45.9	52.8	77.0	90.7
3	Doubao-pro-1.5 🎉	Reason 🧠	<a href="#">Link</a>	2025-01-22	79.5	77.7	88.9	86.0	82.3	62.0	89.7	73.5	88.3	35.1	55.3	77.1	86.5
4	Ovis2_34B	LMM 📈	<a href="#">Link</a>	2025-02-10	77.1	75.5	84.6	89.2	72.8	62.0	80.1	72.2	82.0	37.8	54.9	73.8	85.0
5	Kimi-k1.5	Reason 🧠	<a href="#">Link</a>	2025-01-22	74.9	-	-	-	-	-	-	-	-	-	-	-	-
6	OpenAI o1	Reason 🧠	<a href="#">Link</a>	2024-09-12	73.9	-	-	-	-	-	-	-	-	-	-	-	-
7	Llama 4 Maverick	LMM 📈	<a href="#">Link</a>	2025-04-05	73.7	-	-	-	-	-	-	-	-	-	-	-	-
8	Vision-R1-7B	Reason 🧠	<a href="#">Link</a>	2025-03-09	73.2	75.1	82.7	80.1	70.3	54.7	78.6	65.2	79.1	40.5	45.1	68.9	83.1
9	Gemini 2.0 Flash	LMM 📈	<a href="#">Link</a>	2025-02-05	73.1	-	-	-	-	-	-	-	-	-	-	-	-
10	QVQ-72B-Preview	LMM 📈	<a href="#">Link</a>	2024-12-24	71.4	-	-	-	-	-	-	-	-	-	-	-	-

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

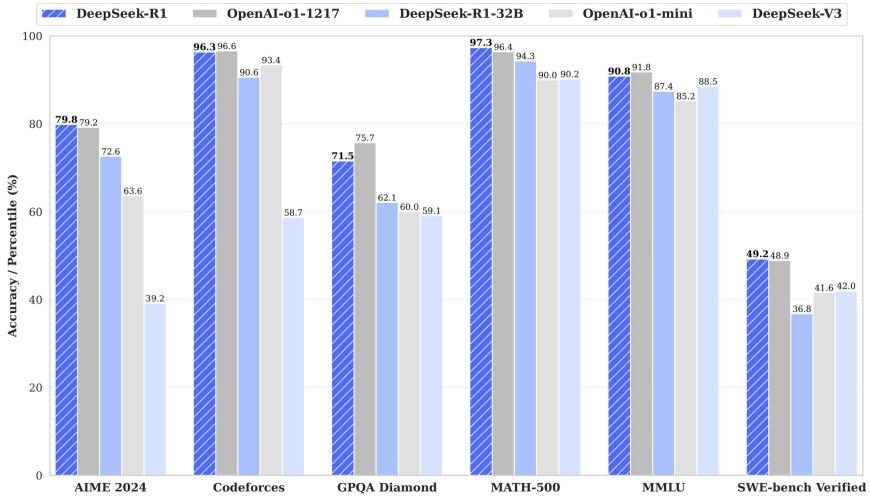


(a) Mathematical reasoning



(b) Visual context

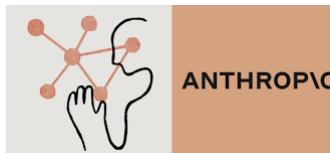
# Recent Released LLMs



	Claude Opus 4	Claude Sonnet 4	Claude Sonnet 3.7	OpenAI o3	OpenAI GPT-4.1	Gemini 2.5 Pro Preview (05-06)
Agentic coding <i>SWE-bench Verified<sup>4,5</sup></i>	<b>72.5% / 79.4%</b>	<b>72.7% / 80.2%</b>	<b>62.3% / 70.3%</b>	<b>69.1%</b>	<b>54.6%</b>	<b>63.2%</b>
Agentic terminal coding <i>Terminal-bench<sup>2,5</sup></i>	<b>43.2% / 50.0%</b>	<b>35.5% / 41.3%</b>	<b>35.2%</b>	<b>30.2%</b>	<b>30.3%</b>	<b>25.3%</b>
Graduate-level reasoning <i>GPQA Diamond<sup>6</sup></i>	<b>79.6% / 83.3%</b>	<b>75.4% / 83.8%</b>	<b>78.2%</b>	<b>83.3%</b>	<b>66.3%</b>	<b>83.0%</b>
Agentic tool use <i>TAU-bench</i>	Retail <b>81.4%</b> Airline <b>59.6%</b>	Retail <b>80.5%</b> Airline <b>60.0%</b>	Retail <b>81.2%</b> Airline <b>58.4%</b>	Retail <b>70.4%</b> Airline <b>52.0%</b>	Retail <b>68.0%</b> Airline <b>49.4%</b>	—
Multilingual Q&A <i>MMMLU<sup>3</sup></i>	<b>88.8%</b>	<b>86.5%</b>	<b>85.9%</b>	<b>88.8%</b>	<b>83.7%</b>	—
Visual reasoning <i>MMMU (validation)</i>	<b>76.5%</b>	<b>74.4%</b>	<b>75.0%</b>	<b>82.9%</b>	<b>74.8%</b>	<b>79.6%</b>
High school math competition <i>AIME 2025<sup>4,5</sup></i>	<b>75.5% / 90.0%</b>	<b>70.5% / 85.0%</b>	<b>54.8%</b>	<b>88.9%</b>	—	<b>83.0%</b>



R1



Claude 4

# A “Simple” Test Question



PL

How many “r” are in “straberrrrry”?

# Struggling to solve “easy” problems



PL How many “r” are in “straberrrry”?

Looking at the word "straberrrry", I can count the letter "r":

s-t-r-a-b-e-r-r-r-r-y

There are **4** letter "r"s in "straberrrry".

Claude Sonnet 4 (May 28, 24)

Gemini 2.5 Pro Preview (May 28, 24)

How many “r” are in “straberrrry”?

Thoughts (experimental)

Expand to view model thoughts

Let's count them:

s-t-r-a-b-e-r-r-r-y

There are **four** "r"s in "straberrrry".

My analysis has narrowed down to two methods: manual scanning and a Python solution. The manual approach yielded a count of four "r"s in "straberrrry". To verify, I constructed a Python script using the `.count()` method, which also confirmed four occurrences. I'm now certain of the solution and ready to articulate it to the user.

Collapse to hide model thoughts

# How Good are Current Advanced Models?



How many baseballs are there?



Let's count the baseballs in each bucket:

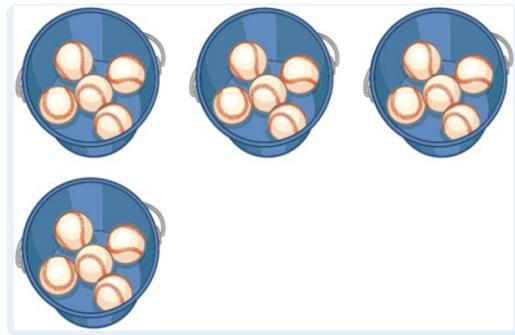
1. First bucket: 5 baseballs
2. Second bucket: 6 baseballs
3. Third bucket: 6 baseballs
4. Fourth bucket: 6 baseballs

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

Answer: 23.



GPT-4V, 2024



There are a total of 19 baseballs across the four buckets:

- The first bucket (top left) contains 4 baseballs.
- The second bucket (top right) contains 5 baseballs.
- The third bucket (bottom left) contains 4 baseballs.
- The fourth bucket (bottom right) contains 6 baseballs.

Adding them together:  $4 + 5 + 4 + 6 = 19$  baseballs.



GPT-4o, Aug 13, 2024

# How Good are Current Advanced Models?



**Question:** What is the highest amount this glass measures?

**Ground truth:** 400 ml

*In the labs, all instruments have the conditions and specifications where they can be used **safely** and **precisely**.*

## 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.

✗

## OpenAI GPT-4o

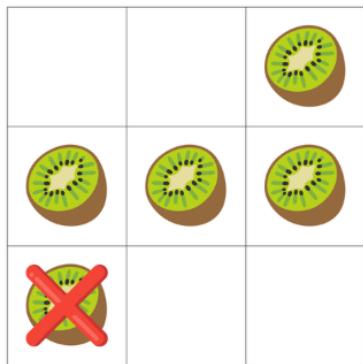
The highest amount this glass beaker measures is 600 milliliters (ml). The markings on the beaker go up to 600 ml, which indicates its maximum volume measurement.

✗

# Struggling to solve “easy” problems



## The Query



**Image path:** rotting\_kiwi.jpg

**Question:** You are given a  $3 \times 3$  grid in which each cell can contain either no kiwi, one fresh kiwi, or one rotten kiwi. Every minute, any fresh kiwi that is 4-directionally adjacent to a rotten kiwi also becomes rotten. What is the minimum number of minutes that must elapse until no cell has a fresh kiwi?

**Ground truth:** 4 minutes

## Wrong Prediction from GPT-4o

To solve this, we need to track how the rot spreads from the rotten kiwi to the fresh ones:

- Minute 1:** The rotten kiwi in the bottom left corner will rot the fresh kiwi directly above it (middle left cell).
- Minute 2:** The newly rotten kiwi in the middle left cell will rot the fresh kiwi directly above it (top left cell) and the fresh kiwi to its right (middle center cell).
- Minute 3:** The rotten kiwi in the middle center cell will rot the fresh kiwi directly above it (top center cell).

After 3 minutes, all fresh kiwis will be rotten.

# Future Directions



# Disentangling Perception and Reasoning



Descriptive Information



Implicit Property



Essential Condition



Text Dominant

Text Lite

Text Only

Vision Intensive

Vision Dominant

Vision Only



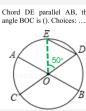
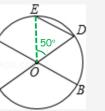
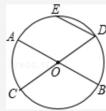
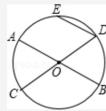
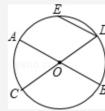
AB and CD are two diameters of circle O, chord DE parallel AB, arc DE is the arc of 50°, then angle BOC is ( ). Choices: ...

Chord DE parallel AB, arc DE is the arc of 50°, then angle BOC is ( ). Choices: ...

AB and CD are two diameters of circle O, chord DE parallel AB, arc DE is the arc of 50°, then angle BOC is ( ). Choices: ...

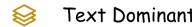
Arc DE is the arc of 50°, then angle BOC is ( ). Choices: ...

Chord DE parallel AB, then angle BOC is ( ). Choices: ...



Text Dominant

Vision Dominant



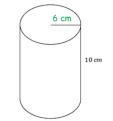
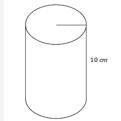
Text Dominant

Vision Dominant

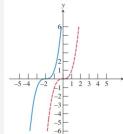
Find the surface area of the cylinder shown. The height is 10 cm and the radius is 6 cm. Give your answer to two decimal places.

Find the surface area of the cylinder shown. and the radius is 6 cm. Give your answer to two decimal places.

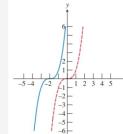
Find the surface area of the cylinder shown. Give your answer to two decimal places.



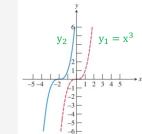
The graph shows  $y_1 = x^3$  passing (0,0) and a vertical or horizontal translation  $y_2$  passing (-2,0). Write an equation for  $y_2$  as shown.



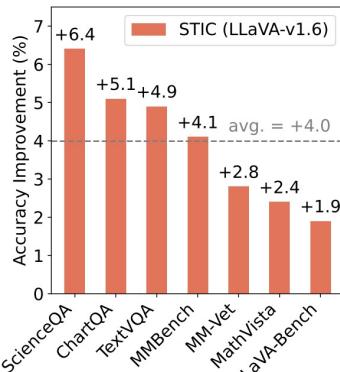
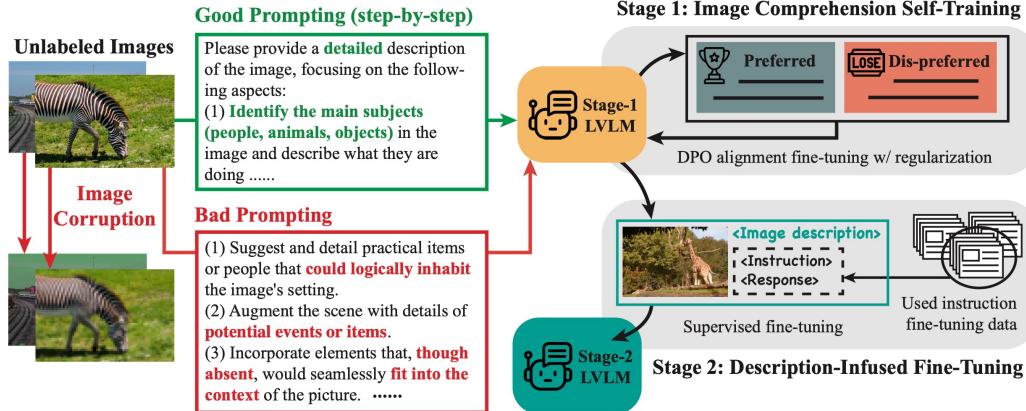
The graph shows  $y_1 = x^3$  and a vertical or horizontal translation  $y_2$ . Write an equation for  $y_2$  as shown.



The graph shows  $y_1$  and a vertical or horizontal translation  $y_2$ . Write an equation for  $y_2$  as shown.



# Self-Training VLMs on Image Comprehension



**Query:** How many gallons of supreme gasoline can I get with \$50?

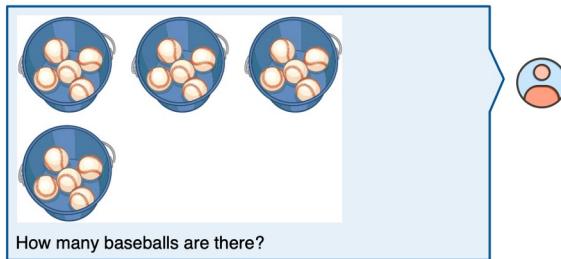
**Base (LLaVA-v1.6 7B):**

Based on the current gas prices displayed on the sign, you can get **approximately 3.65 gallons** of supreme gasoline with \$50.

**STIC (LLaVA-v1.6 7B):**

With \$50, you can get **approximately 13.69 gallons** of supreme gasoline, as indicated by the price of **\$3.65 per gallon** on the sign.

# Optimizing Responses via Feedback



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:

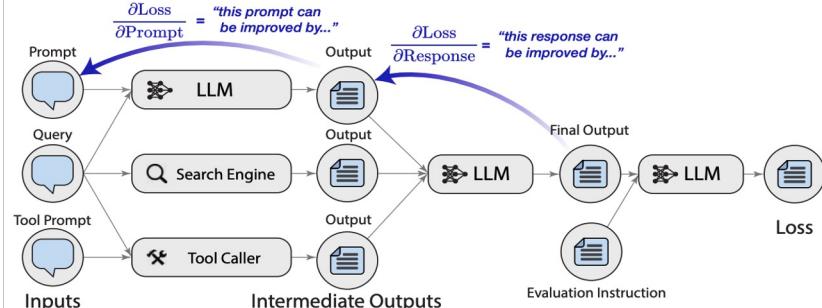
1. First bucket: 5 baseballs
2. Second bucket: 5 baseballs
3. Third bucket: 5 baseballs
4. Fourth bucket: 5 baseballs

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



Answer: 20.

## b Blackbox AI systems and backpropagation using natural language 'gradients'



nature

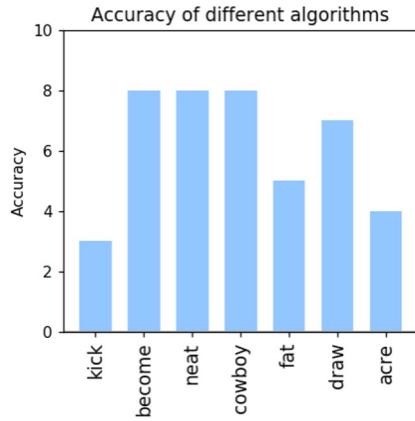
Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [articles](#) > [article](#)

Article | Published: 19 March 2025

## Optimizing generative AI by backpropagating language model feedback

# TextGrad for Multimodal Math Reasoning



**Question:** What is the sum of the accuracies of the algorithms fat and acre?

**Prediction:** The accuracy of "fat" is represented by the fourth bar from the left, which shows an accuracy of 6. The accuracy of "acre" is represented by the last bar on the right, which shows an accuracy of 4. Now, we add these two values together:  $6 + 4 = 10$ . The correct sum of the accuracies of the algorithms "fat" and "acre" is 10.

**Prediction:** The accuracy of "fat" is represented by the fifth bar from the left, which shows an accuracy of 5. The accuracy of "acre" is represented by the last bar on the right, which shows an accuracy of 4. Now, we add these two values together:  $5 + 4 = 9$ . The correct sum of the accuracies of the algorithms "fat" and "acre" is 9.

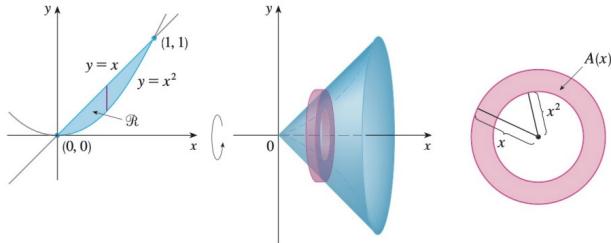
 **TextGrad**

 **Gradients**

The existing answer provides a clear and logical approach to solving the problem by interpreting the bar chart and summing the accuracies of the specified algorithms. However, there is an error in the interpretation of the values from the chart. Upon reviewing the bar chart: the accuracy of "fat" is represented by the fifth bar from the left, which shows an accuracy of 5.

**Problem**

The region  $\mathcal{R}$  enclosed by the curves  $y = x$  and  $y = x^2$  is rotated about the  $x$ -axis. Find the volume of the resulting solid.

**Image****Correct Solution**

The curves  $y = x$  and  $y = x^2$  intersect at the points  $(0, 0)$  and  $(1, 1)$ . The region between them, the solid of rotation, and a cross-section perpendicular to the  $x$ -axis are shown in the Figure. A cross-section in the plane  $P_x$  has the shape of a washer (an annular ring) with inner radius  $x^2$  and outer radius  $x$ , so we find the cross-sectional area by subtracting the area of the inner circle from the area of the outer circle:

$$A(x) = \pi x^2 - \pi (x^2)^2 = \pi (x^2 - x^4)$$

Therefore we have

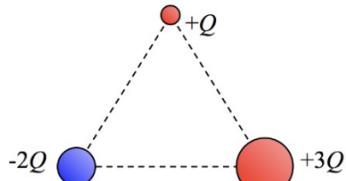
$$\begin{aligned} V &= \int_0^1 A(x) dx = \int_0^1 \pi (x^2 - x^4) dx \\ &= \pi \left[ \frac{x^3}{3} - \frac{x^5}{5} \right]_0^1 = \frac{2\pi}{15} \end{aligned}$$

**Final Answer:**  $\frac{2\pi}{15}$

# College-Level Scientific Problem-Solving



## Multimodal reasoning question in EMMA



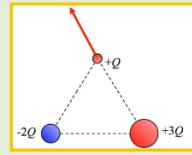
**Question:** Three point charges, of charge  $+Q$ ,  $-2Q$ , and  $+3Q$ , are placed equidistant as shown. Which vector best describes the net direction of the electric force acting on the  $+Q$  charge?

- a.
- b.
- c.
- d.
- e.

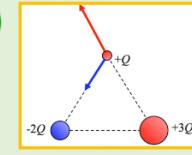
## Human reasoning

The direction of the electric force due to  $-2Q$  and  $+3Q$  can be visualized with a quick free-body sketch of the electric forces:

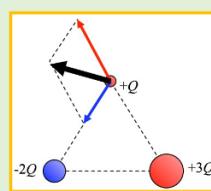
1



2



3



Combining the forces:  
∴  $|+3Q| > |-2Q|$   
∴ **repel** > **attract**

## Model reasoning (GPT-4o)

### 1. Forces Due to the Charges:

- The  $-2Q$  charge attracts the  $+Q$  charge.  
(Force points downward-left).
- The  $+3Q$  charge repels the  $+Q$  charge.  
(Force points downward-right).

### 2. Force Magnitudes (Coulomb's Law):

- Force due to  $-2Q$ :  $F_{-2Q} = \frac{2kQ^2}{r^2}$ .
- Force due to  $+3Q$ :  $F_{+3Q} = \frac{3kQ^2}{r^2}$ .  
.....

**Error:** While the model understands that like charges repel, it incorrectly identifies the force from  $+3Q$  to  $+Q$  as downward-right ↘, when it should be upward-left ↗.

- 🔎 **Scope:** multimodal math reasoning involves diverse **reasoning** and **visual** types.
- 📜 **Three recipes:** **multimodalities**, **knowledge**, and **rigorous reasoning**.
- 📁 **Key datasets:** individual areas or **comprehensive** evaluations.
- ⚒ **Evaluation:** answer-based, **LLM-as-Judge**
- ✕ **Approaches benchmarked:** VLMs, neuro-symbolic, tool-augmented, agentic models
- 📈 **SOTA & challenges:** may achieve **high scores**; still **fail to “simple” problems**
- 💎 **Future directions:** more robust benchmarks along with effective methods to enhance **both perception and reasoning** capabilities